Summer 8-15-2017

Sequence Determinants of the Individual and Collective Behaviour of Intrinsically Disordered Proteins

Alexander S. Holehouse
Washington University in St. Louis

WASHINGTON UNIVERSITY IN ST. LOUIS

Division of Biology and Biomedical Sciences

Computational and Molecular Biophysics

Dissertation Examination Committee:
Rohit V. Pappu, Chair
Jan Bieschke
Gregory R. Bowman
Clifford P. Brangwynne
Susan K. Dutcher
Robert Mecham

Sequence Determinants of the Individual and Collective Behaviour of Intrinsically Disordered Proteins

by

Alexander Steven Holehouse

A dissertation presented to
The Graduate School
of Washington University in
partial fulfillment of the
requirements for the degree
of Doctor of Philosophy

August 2017
Saint Louis, Missouri

Contents

List of Tables

List of Figures

Acknowledgments

In the summer 1996 I was watching the reboot of the animated series 'Flash Gordon', and was - for whatever reason - rather impressed by Dr. Hans Zarkov. The other characters kept calling him 'Doc', so as an impressionable young man I decided that I too would change my name to 'Doc'. To my dismay, I was sternly informed that I had to have a Ph.D. to be called 'Doc', and so, here we are.

There are far too many people I need to thank. First and foremost I need to thank the head honcho, Prof. Rohit V. Pappu. I can honestly say that working with Rohit for the last four years has been the greatest privilege of my life. It's tempting to think that his "*relentless pursuit of excellence*" is a lofty but perhaps overambitious goal, yet day after day his ability, drive, and determination prove that not only are such heights achievable, but that you can have fun along the way. His vision and direction coupled with an unnerving ability to name the page number of the paper in which the equation you need, equips him with an ability to synthesize and distil ideas that is beyond anyone I've ever known. The intellectual freedom he has given me has been empowering, and has allowed me to develop as a scientist in a way I did not anticipate during a Ph.D., yet I also know when I get stuck he will at *worst* know how to get unstuck, and at best have already solved *that exact problem*[1]. I could go on, but suffice to say, I am very much looking forward to continuing our work together.

[1]This has happened multiple times

Secondly, I would like to thank the other members of the lab. In particular, Ammon, Anu, Kiersten, Megan, Nick and Tyler, and Rahul have played pivotal roles throughout my development as a scientist in a wide range of capacities. Their guidance, advice, and distractions have been welcome and necessary, and much of this work is coated in their fingerprints. In addition it has been a pleasure to work with James, Jared, Jeong-Mo, Kanchan, Mary, Max, Martin, Minerva, and Sang-Eun; watching the lab evolve has been a joy, and I am excited to see where the next generation of members takes us. I would also like to extend a special thank you to Andreas for all his help in many different capacities.

This thesis would not be what it is without the exceptional collaborators I have had the good fortune of working with over the last four years. In particular, working with Erik and Tanja (chapter 6), Ivan, Dan, and Osman (chapter 7), Steven, Shani and Cliff (chapters 12 and 13), and Chi and Mike (chapter 11), Kristen (chapter 4), Sebastian König, Ben Schuler, and Andrea Soranno, Thomas Boothby and Gary Pielak, Titus Franzmann and Simon Alberti, Sam Powers, Katie Schreiber, and Lucia Strader, Jamie Allen and Gregg Jedd, Max Staller and Barak Cohen, Jingyi Fei, Linda Pike, Petra Levin, Jan Bieschke, John Cooper, Vasilios Kalas, and last but by *no* means least, Alex Barrow. Each one of these collaborators has forced me to think deeply about challenging and interesting problems and allowed me to build a broad intellectual toolkit that, I hope, will allow me to assess biological questions via the context in which they appear. I cannot thank you all enough. Thank you to my thesis committee (and especially Greg Bowman) for continued support, guidance, and advice on a wide range of topics.

I would also like to thank my friends. Special "shout-out" to Will, Mariah, Kelly, Apurwa, Lin, Yerdos, Anne, Robb, and Josh for many years of food, drink and fun. I started listing the various people I have spent an inordinate amount of time with at bars, brunch, restaurants,

and on Frisbee fields, but the list rapidly became too long. From drinkwalkin' to ordering many times more food than we could possibly eat to righting the worlds wrongs as the bar-staff tried to quietly usher us out, you know who you are, and you have played integral role in this. A special thanks needs to go to the wonderful pARCHd ultimate Frisbee team, a genuinely exceptional collection of people who provided a weekly highlight for me that was *mostly* outside of science (who could ignore our ground-breaking work on the science of dynamic planks[2]).

Thank you to Kent and Bonnie Lattig and the Center for Biological Systems Engineering for generous funding, the National Science Foundation, National Institute of Health and the Division of Biological and Biomedical Sciences (DBBS) for much of my graduate funding and to Kayak's Coffee and Northwest Coffee Roasting for months of free WI-FI, excellent coffee, and the space to think.

Finally I need to thank my family. Anne and Bill have been like parents to me, accepting me as their son-in-law with open arms and providing a home-from-home over many years, including a wonderful extended family in Marnie & Bob and the rest of the Mueller family. Their kindness knows no bounds, and having an ally to irritate Martha with is a real joy for everyone (except Martha). My parents have also been like parents to me, but as I've gotten older it has become increasingly clear that they are simply not normal people; their kindness, empathy, and fairness coupled with a true drive is a source of constant and sometimes intimidating inspiration. My brother, Rob, remains may favourite anecdotal example of either "following one's dream's" or "blind stubbornness" (I suspect the former), and his success and relentless creativity fills me with overwhelming pride. A finally I need to thank Martha (aka Dr. Bucky). Despite my best efforts to bore her to death, she has at least

[2]Gray, Hautly, *et al.* (unpublished)

remained awake through many hours of me explaining my work to her. She makes everything worthwhile; none of this matters even a little without her, and I *remain* excited about our lifetime of adventures together.

Alexander Steven Holehouse

Washington University in Saint Louis
August 2017

Dedicated to Bucky

ABSTRACT OF THE DISSERTATION

Sequence Determinants of the Individual and Collective Behaviour of Intrinsically Disordered Proteins

by

Alexander Steven Holehouse

Doctor of Philosophy in Computational and Molecular Biophysics

Washington University in St. Louis, August 2017

Rohit V. Pappu

Intrinsically disordered proteins and protein regions (IDPs) represent around thirty percent of the eukaryotic proteome. IDPs do not fold into a set three dimensional structure, but instead exist in an ensemble of inter-converting states. Despite being disordered, IDPs are decidedly not random; well-defined - albeit transient - local and long-range interactions give rise to an ensemble with distinct statistical biases over many length-scales. Among a variety of cellular roles, IDPs drive and modulate the formation of phase separated intracellular condensates, non-stoichiometric assemblies of protein and nucleic acid that serve many functions. In this work, we have explored how the amino acid sequence of IDPs determines their conformational behaviour, and how sequence and single chain behaviour influence their collective behaviour in the context of phase separation.

In part I, in a series of studies, we used simulation, theory, and statistical analysis coupled with a wide range of experimental approaches to uncover novel rules that further explore

how primary sequence and local structure influence the global and local behaviour of disordered proteins, with direct implications for protein function and evolution. We found that amino acid sidechains counteract the intrinsic collapse of the peptide backbone, priming the backbone for interaction and providing a fully reconciliatory explanation for the mechanism of action associated with the denaturants urea and GdmCl. We discovered that proline can engender a conformational buffering effect in IDPs to counteract standard electrostatic effects, and that the patterning those proline residues can be a crucial determinant of the conformational ensemble. We developed a series of tools for analysing primary sequences on a proteome wide scale and used them to discover that different organisms can have substantially different average sequence properties. Finally, we determined that for the normally folded protein NTL9, the unfolded state under folding conditions is relatively expanded but has well defined native and non-native structural preferences.

In part II, we identified a novel mode of phase separation in biology, and explored how this could be tuned through sequence design. We discovered that phase separated liquids can be many orders of magnitude more dilute than simple mean-field theories would predict, and developed an analytic framework to explain and understand this phenomenon. Finally, we designed, developed and implemented a novel lattice-based simulation engine (PIMMS) to provide sequence-specific insight into the determinants of conformational behaviour and phase separation. PIMMS allows us to accurately and rapidly generate sequence-specific conformational ensembles and run simulations of hundreds of polymers with the goal of allowing us to systematically elucidate the link between primary sequence of phase separation.

Preface

In the interest of clarity, I felt it would be useful to provide a general outline for the structure of this thesis.

Part I is a general introduction to many of the topics of key relevance to this work. This includes protein biophysics, intrinsically disordered proteins and protein regions (IDPs), and biological phase separation. These chapters can be treated as reviews of the relevant literature, and while do not introduce much novel information, provide a convenient synthesis of the state-of-the-art.

Part II covers the first half of my thesis work; the sequence determinants of the individual behaviour of unfolded proteins. Given a monomeric unfolded protein, how and why does the amino acid sequence determine its solution behaviour.

Part III describes the second half of my thesis work; the sequence determinants of the collective behaviour of IDPs in the context of phase separation and gelation. How are the sequence features associated with disordered proteins coupled to their macroscopic phase behaviour?

The goal in structuring this thesis in this way is twofold. Firstly, I hope to avoid redundancy (within reason), which is of enormous benefit to me and to the reader. Secondly, by ensuring critical ideas and concepts are presented in a coherent and self-contained manner they will hopefully be useful chapters for future lab members. Finally, parts II and III, while highly

related, represent a distinct set of ideas, performed with a distinct set of collaborators and tools.

Finally, there is a single idea I would like people to take away from this work. At the time of writing, there is substantial scientific discussion as to the mechanism and role of various types of physical processes in cellular function. In the context of disordered proteins; to what extent does conformational behaviour effect fitness and function? In the context of intracellular condensates, are they liquid or are they solid, are they formed by cation-pi interactions or are they formed by β-sheet mediated interactions, are the folded domains driving self assembly and specificity, or is this facilitated by RNA?

The answer to all these questions is "yes".

We wish to understand mechanism through the elucidation of design principles, yet evolution does not select for principles, it selects for fitness, an epistatic and emergent property. If similar outcomes can be achieved in different but equivalently fit ways, then given the stochastic nature of evolution this is almost guaranteed to happen. We have specific examples where every statement in the preceding paragraph is true. We do not need one person to be right or wrong; our nascent understanding of complex biological systems is that the space of information-processing solutions is astronomical. Think of the diversity observed in structural biology - the repertoire of tertiary structures is enormous. There are countless examples of nearly identical functions being performed by proteins with radically different structure. This divergence, this variety in structure and function, is what makes evolution robust. It is an inherent bet-hedging mechanism woven into the fabric of statistical physics. On the contrary, the desire to categorise and abstract complexity into distinct groups is an inherently human endeavour. Much as we may wish and as convenient as it would be, Nature does not have a plan.

Part I

Background

Chapter 1

Fundamentals of Protein Biophysics

Proteins are important. From the connective tissue that binds our bodies together, to the signalling machinery that converts photons into chemical information, proteins are involved in a significant fraction (and indeed, perhaps the majority) of the 'interesting' biology that occurs in nature[3] [5]. Since it was first suggested and later shown that proteins are the ubiquitous components that mediate cellular function, understanding how these biological macromolecules are able to perform such a wide variety of tasks has been a central goal of biology [5, 35, 343]. The discovery that many proteins exist in well defined yet non-symmetrical 3D structures won Perutz and Kendrew a Nobel prize, and the elucidation and application of the structure-function paradigm has perhaps been the most powerful single idea in modern biochemistry and biophysics [104, 210, 285, 322, 365, 408, 415, 484, 500].

The structure-function paradigm is based on a simple yet profound idea. Many proteins exist as conformationally well defined molecular machines. Their function depends on this structure. Consequently, structure determines function, and often this function relies on a mechanical change in the protein's structure. Complex allosteric networks allow information

[3]While we present a protein-centric description of biology, we should not forget about our friends the nucleic acids or carbohydrate based species, several of which we will return to in later chapters

to flow through the interior of a protein, flexible hinges allow substrates to enter and exit enzymes, and exquisitely tuned pores provide selectivity for ions as they cross membranes [151,229,549]. Moreover, the impact of a genetic mutation - a hereditary or somatic changes that cause amino acids in the protein sequence to be added, deleted, or entirely changed - is rationalisable based on the structural impact that mutation may have [629]. Yet, despite the power of the structure-function paradigm, in the last thirty years or so it has been become apparent that the structure-function relationship is more complicated than first anticipated [652].

Many proteins contain regions that unable to fold autonomously, and instead exist in a floppy, unstructured, or disordered ensemble of inter-converting states [608]. Initially dismissed as an artefact of *in vitro* preparation techniques, it has become clear that these intrinsically disordered proteins and protein regions (collectively referred to as IDPs) exist *in vivo* and are functionally relevant [158,587,588,603,605,608,654]. Furthermore, while it was originally believed that these regions would serve as simple flexible linkers - or perhaps at most entropic springs - an explosion of research over the last ten years has illustrated how disorder performs a myriad of functions, from molecular recognition to dynamic self-assembly [67,606,654]. In analogy to the structure-function paradigm proposed previously, IDPs show an ensemble-function relationship.

The body of this thesis is divided into three sections. The introductory section contains three introductory chapters describing protein biophysics, intrinsically disordered proteins, and phase separation in biology. The next section (chapters 4 - 10) describes our work on the sequence determinants of individual disordered proteins. The final section (chapters 11 - 15) describes our work on the sequence determinants of the collective behaviour of disordered proteins in the context of biological phase separation. Chapter 1 has been included to ensure

the prerequisite content is appropriately covered, given so much of this work focuses on the fundamental relationship between sequence and conformation.

1.1 A Hierarchical Description of Protein Structure

Proteins are heteropolymers of amino acids. This is a simple, but profound concept that we should spend some time unpacking. A polymer consists of many base units (monomers) connected in series to form a long chain-like molecule [504]. Homopolymers consist of only one repeating monomer. Heteropolymers (such as proteins) contain a mix of different monomers; in the case of proteins those different monomers are the amino acids. There are twenty natural amino acids, each of which has a unique sidechain that imparts a distinct set of of molecular features (for an overview of the natural amino acids see appendix C) [35]. Moreover, many of the amino acids can undergo post-translational modifications (PTMs), which involve the reversible ligation of a new chemical groups onto specific sidechains [358]. These PTMs provide a mechanism for biology to dynamically regulate the chemical composition of polypeptides in response to distinct signals, allowing the cellular state to be written, read, and erased in an rapid and controllable fashion [278]. With twenty distinct chemical building blocks and no major constraints on the order in which they appear, Nature has an enormous tool-kit with which it can construct incredibly complex heteropolymers, which in turn can undergo chemical editing via PTMs.

Many proteins undergo an autonomous (or semi-autonomous) re-arrangement to fold into a well defined three dimensional structure [143]. This process of **protein folding** represents one of the most well studied phenomena in biophysics. For many proteins this folded state is synonymous with their native state - both *in vivo* and *in vitro* many proteins will

robustly fold into and then remain in their protein-specific folded state for months or even years. This folded state is also typically associated with the protein's cellular function, be it catalysis, molecular recognition, or structural integrity [35]. How does this folded structure arise? In the following subsections we will briefly overview the five levels of protein structural organization. We will then consider protein dynamics, the thermodynamic origins of protein folding, and finally discuss a range of putative models for describing the mechanism(s) through which proteins fold.

1.1.1 Primary Structure

The 'primary structure' (or 'primary sequence') of a protein refers to the specific order in which the amino acids appear in the peptide sequence [35]. Conveniently, this can be written as a linear sequence of letters. In terms of protein structure, this is the only information explicitly encoded by our genome. The three-dimensional state of a protein is an emergent property of the amino acid sequence, the solvent environment (including binding partners), and the physical properties of the system (temperature, pressure *etc.*).

Given that the primary sequence can be directly determined from the genome, and thousands of organisms have now had their genomes fully sequenced, on a very practical level obtaining protein sequence information is now trivial and instantaneous [600]. Can we use this sequence information to inform on the expected structural and functional properties of a protein? As will be discussed in the following subsections, a wealth of structural and functional data has allowed the construction of large databases that in turn can be used to generate predictive models that can classify and annotate novel sequences [176,210,408,500]. Beyond informatics

based approaches, physics based models provide a means to extract three-dimensional information from primary sequence alone [64,349,401,433,474,495,650]. While *de novo* structure prediction remains challenging, a large body of work from several labs has provided a general framework through which *de novo* structure prediction can be done, to some degree of accuracy. The recent development of Bayesian-based methods represent a promising new approach for combining physics-based models with sparse experimental data [349, 447]. For the majority of common organisms, complete and annotated proteomes exist, and new proteomes are being added on a near weekly basis [600]. As a result, we have an extraordinary wealth of primary sequence data, and approaches that are able to extract real, novel insight from this data alone are an extremely valuable.

1.1.2 Secondary Structure

Secondary structure is often considered to be the local structural units through which folded proteins are assembled. It represents a set of common structural motifs that are reused throughout the kingdoms of life [274, 464]. While different amino acids have strong preferences for and against certain types of secondary structure, the energetic driving force for secondary structure originates from hydrogen bonding in the protein backbone [112]. As a result, secondary structure is often repetitive, and while sidechain position varies from protein to protein, the backbone configuration is, broadly speaking, well defined by the backbone ϕ and ψ angles (the dihedral angles associated with the amino acid backbone, see appendix C). Because of this well defined backbone behaviour, the Ramachandran map - which describes a 2D space defined by these ϕ and ψ angles - facilitates a simple method to assign the secondary structure state of each amino acid in a folded protein structure [35,486].

There are relatively few distinct types of secondary structure. The most common are the α-helix and the β-strand/β-sheet [274,484]. These structures are shown below in fig. 1.1. In addition to the α-helix and the β-strand there are various other structural motifs such as the the PPII helix and then 310-helix. Beyond these, various other less common structural motifs exist, including loops and turns are sometimes considered *bona fide* secondary structure elements [323, 485].

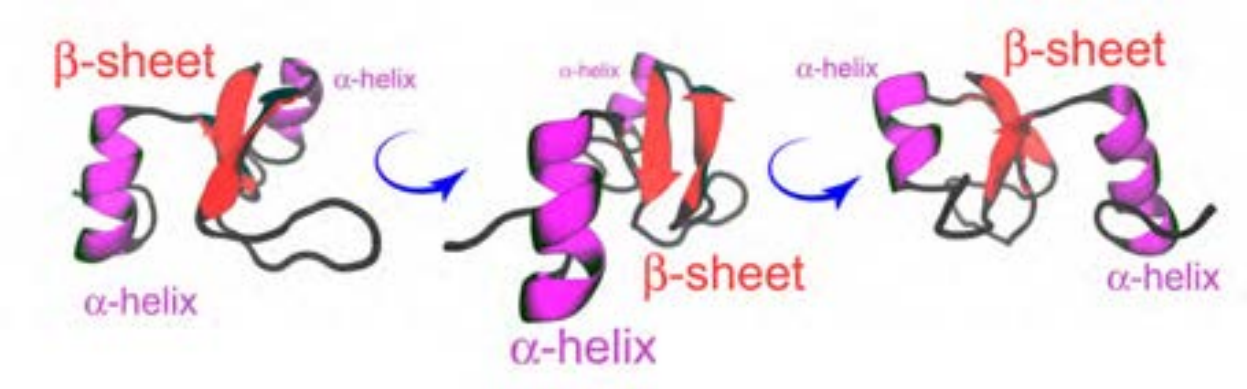

Figure 1.1: Secondary structure elements (α-helices and β-sheets) shown for the protein NTL9. The β-sheets are made up of two β-strands.

Secondary structure frequently gives rise to the core structural regions in folded proteins, and is often considered a structural building block upon which tertiary and quaternary structure is assembled [484, 486]. For helices in particular, the driving forces for formation are often (though not always) fairly local, meaning they can form early during protein folding. Several models for protein folding suggests that for the nucleation of helices is a typically early step in the folding process due to their locally cooperative nature of formation, although this need not necessarily be the case [642].

1.1.3 Tertiary Structure

Tertiary structure is typically what most structural biologists think of when they consider protein structure [486]. It refers to the three-dimensional arrangement of chains, helices, sheets and loops that give rise to a globular, folded protein. Tertiary structure refers to the folded conformation adopted by a single polypeptide chain. Before the first structure (myoglobin) was determined, there had been a quiet expectation that protein structures would be semi-crystalline and symmetrically ordered. Upon the revelation that myoglobin was an asymmetric, amorphous blob, Max Perutz was unable to hide his distaste [285]:

> *"Could the search of ultimate truth really have revealed so hideous and visceral-looking an object?"*

Irrespective of Perutz' disgust, tertiary structure is typically the level at which structural information can provide direct mechanistic insight. The relative position of secondary structure elements and intervening loops coupled with the chemical composition of the residues positioned along those elements provides the structural scaffold that gives rise to function [5, 343, 484]. In general, single folded domains are typically on the order of 40-200 residues in length [334]. More complex proteins are composed of multiple folded domains that are in direct interaction with one another, or are connected by flexible linkers [330, 414]. The Pfam database characterize the types of domains observed in nature, as identified by hidden Markov Models, and represents the gold standard in terms of unpacking the three-dimensional structural units associated with an amino acid sequence [176].

To obtain three-dimensional information on a protein (hereafter referred to simply as 'structural information') we require techniques that provide insight into the relative positions of

atoms. X-ray crystallography and NMR have been the primary methods for obtaining full structural descriptions of proteins [90, 482, 530]. These methods utilize electromagnetic radiation (X-rays or radio-waves) to construct a self-consistent three-dimensional model of the relative position of atoms in a protein. X-ray crystallography relies on the fact that proteins can form homogeneous crystals when combined with various organic and inorganic solvents and driven to high concentrations of protein. These crystals are effectively an ordered three-dimensional lattice, and the electron-dense regions associated with the lattice will diffract X-ray beams to create a characteristic diffraction pattern. This diffraction pattern can be converted into a three-dimensional electron density map, which can be used to elucidate the structure of the protein.

NMR uses an entirely different approach; instead of diffracting low-wavelength electromagnetic ration, it uses radiowaves to drive the transfer of magnetization between certain nuclei (through space or through bonds), allowing the NMR spectroscopist to construct a set of positional restraints which in combination with the primary sequence can be used to obtain a family of structures that are consistent with the available data. While NMR-based structure determination is *possible*, it is typically significantly more challenging than crystallography and is generally only used in cases where crystallography has failed. There is also an upper size limit of around 100 kDa imposed by the inherent slower rotational diffusion ('tumbling') of larger proteins.

Despite the fact that NMR and crystallography are fundamentally different in terms of approach, underlying assumptions, and associated solution conditions, protein structures determined by both methods independently are remarkably similar [544]. The interpretation of this result is that although the crystal structure may be susceptible to some crystallization artefacts, broadly speaking the conformational state observed from X-ray crystallography is

highly similar to at least one of the possible solution state structures. For a sense of scale, at the time of writing, there are $115,267$ X-ray crystallography structures in the protein data bank (PDB) and $11,759$ NMR structures, while ten years ago there were under $\sim 50,000$ structures combined [36]. We mention this, only to make the point that we are truly in a golden age of structural data.

In the last few years, cryo-electron microscopy (cryoEM) has rapidly emerged a new approach for obtaining atomic-resolution structural insight into tertiary (and quaternary) structure [24]. While further discussion on cryoEM is entirely beyond the scope of this work, it seems important to mention this technique, given all indications are that cryoEM will become the norm for structural determination of larger protein complexes going forward.

Beyond these methods to perform complete structural determination, there are a host of methods for assessing specific structural features. These include fluorescence based approaches such as Förester Resonance Energy Transfer (FRET), which use the extent of non-radiative energy transfer between two exogenously placed dye molecules to infer a distance distribution (FRET will be discussed further in chapters 2 and 7) [145]. Similarly, EPR and PRE based methods provide specific distance profiles using NMR based methodologies, which in conjunction with computational models provide a new route for *de novo* structure determination [349]. Contact quenching methods provide information of pairwise dynamics which can be interpreted as structural insights via additional constraints and assumptions [317].

Finally, a set of methods provide information of the global dimensions of proteins. These includes scattering techniques such as small angle X-ray scattering (SAXS), static light scattering (SLS), dynamic light scattering (DLS), Fluorescence correlation spectroscopy (FCS)

and various other techniques [253, 306, 467]. FCS will be discussed further in chapters 5 and 12, and SAXS in chapters 6 and 7.

1.1.4 Quaternary Structure

Many proteins assemble into multi-subunit complexes consisting of several independently synthesized polypeptide chains that interlock and interact with one another to form a large macromolecular assembly with well defined stoichiometry. The relative orientation of the chains and the resulting assembly is referred to as quaternary structure. These complexes may be heteromeric, made up of multiple different proteins (e.g. the F_1F_0-ATPase, or homomeric, such as homodimers or homoteteramers (e.g. haemoglobin) [2, 545]. Typically determining such quaternary structure relies on methods such as X-ray crystallography or cryoEM. Other methods, including cross-linking followed by mass spectrometry provide insight into the specific regions that engage in intermolecular interactions [488].

1.1.5 Quinary Structure

Unlike primary - quaternary structure, which describe precise arrangements of either sequence (primary) or structure (secondary - quaternary), quinary structure is a much more general term. We use it here to refer to protein-assemblies containing a large number of proteins with a variable stoichiometry [159, 373, 388, 483, 607, 622]. In the original definition by Vainshtein in 1973 the term was coined to describe the "*combination of molecules of proteins, nucleic acids, and nucleoproteins into aggregates*" as visible by electron microscopy [607]. This is an oddly prescient definition, as will become clear in chapter 3. While recent usages have suggested that the interactions that give rise to quinary assemblies must be weak,

nothing in the original definition makes any suggestion regarding the strength of those interactions. We believe this is a useful term as it provides a general framework to describe a wide range of protein assemblies without assuming their underlying thermodynamic or kinetic details.

1.2 A Hierarchical Description of Protein Dynamics

Having introduced key concepts in protein structure, it is relevant to also introduce the equivalent concepts in protein dynamics. This body of work will, in general, focus much less heavily on dynamics and kinetics than it will on thermodynamics, in part because much of the computational work done here has taken advantage of simulations that generate thermodynamic ensembles but do not provide direct insight into the associated kinetic behaviour. However, at least a cursory overview of protein dynamics is crucial for understanding many of the experimental techniques that will be described in the coming chapters.

It is important to recall that even folded proteins are flexible and dynamic macromolecules engaging in conformational fluctuations on a wide range of both length scales and time scales. The fastest motion is that of bond stretching and vibration, which occurs on the 10-100 fs timescale. Experimentally such motions are accessible via infrared (IR) spectroscopy, which provides information on resonance structures and bond characteristics [131]. Lateral motion of atoms occurs on a timescale of $\sim$ 1 ps, and can broadly be considered the 'jiggling' of atoms [519]. Sidechain rotation occurs on the order of $\sim$ 100 ps, although this will be hugely determined by the solvent accessibility of the sidechain and its surrounded chemical environment [105, 536]. For solvent buried sidechains, rotation can be on the order of ms to seconds,

depending on proteins stability and the interactions experienced by the sidechain. Multi-residue reconfigurations of flexible linkers and disordered regions (of say 50-100 residues), an approximate timescale of 20 - 200 ns, depending on the properties of the chain, should be expected [554–556]. Larger-scale conformational changes of folded domains can take anywhere from a few miliseconds to a few minutes [139, 316, 620]. Finally, proline *cis*-to-*trans* isomerization has a high energy barrier of ∼20 kcal/mol giving it a characteristic transition time of 700 - 900 s. Naturally, all these numbers are rough estimates and may not necessarily apply to specific instances. However, they provide some sense of the time scales associated with protein dynamics.

As with protein structure, these dynamics are heavily dependent on the solution environment [554–556]. Even in aqueous solutions, the protein surface will enslave water in a locally retarded hydration shell that shows reduced dynamics when compared to bulk water [423, 424]. A necessary consequence of this result is that the solvent has an impact on protein dynamics, providing a coupling between solution properties and functional kinetics.

1.3 Protein Folding

A key idea postulated by Anfinsen in 1973 is as follows [9]:

> *The three dimensional structure of a native protein in its normal physiological milieu ... is the one in which the Gibbs free energy of the whole system is lowest; that is, that the native conformation is determined by the totality of interatomic interactions and hence by the amino acid sequence, in a given environment.*

Part of the justification for this statement is that the majority of foldable small single domain proteins (< 200 residues) will reversibly fold and unfold *in vitro*. This demonstrates that protein folding to the native state can occur in the absence of any of the cellular proteostatic machinery such as chaperones, stabilizing osmolytes, or an external energy source. Critically, it also shows that the crowded cytosol, a densely-packed soup of millions of different macromolecules, is not a key determinant in the acquisition of structure. For larger proteins or those that are typically not found in the cytosol, such as membrane proteins or extracellular protein, additional components such as chaperones are required to facilitate correct folding and/or cellular localization. However, simple single domain proteins are able to autonomously drive their own self assembly.

In contrast to Anfinsen's postulate, it has been proposed that for many proteins the native state may actually represent a *kinetically* stable state and not the true thermodynamic minimum, especially at high concentrations of protein [26, 194, 296]. When incubated at high concentration, many proteins will undergo a transformation to form long, hyper-stable cross-β amyloid fibers [296]. Amyloid fibers are more commonly associated with disease-linked aggregates, such as a neurofibrillary tangles or amyloid-beta plaques in Azheimer's disease. However, many commonly studied proteins that have no known association with amyloid formation (pathological or otherwise) - such as lysozyme, insulin, myoglobin - have been shown to form amyloid fibrils at high concentration [55, 170, 268]. One possible explanation is that the formation of the amyloid state represents a (largely) sequence-independent thermodynamic minimum associated with high protein concentration that is primarily driven by polypeptide backbone interactions. This hypothesis is in line with a growing body of evidence suggesting that these large, hyper stable amyloids are largely inert, whereas less

stable oligomers and amorphous aggregates may be more strongly associated with toxicity [47]. Further work is needed, both *in vitro* and *in vivo* to fully determine the relative position of the native state on the full free energy diagram.

The ability of small single domain proteins to autonomously folder under conditions that are entirely orthogonal to the cellular environment is (arguably) the primary reason why structural biology has been as prolific and successful as it has been. The corollary of this success is that it has fundamentally defined our collective understanding of what a protein is and how it folds (NB: the entire notion that there is a standard protein may be part of the issue) [93, 143, 147, 198, 254, 496]. To what extent are the results determined from a handful of small stable proteins that express well in *E. coli* generalizable to all proteins? In terms of structure - fairly well. Many of the structural features identified in early crystallographic studies have been identified in proteins sampled from across the kingdoms of life in much more complex structures determined via new approaches. However, it remains to be seen if all of the lessons learned from small, fast-folding proteins will be applicable to the protein folding problem in general. Considering all this what are the fundamental forces that drive protein folding? More specifically, what are the chemical motifs that facilitate protein-protein interaction, be they intramolecular (as in protein-folding) or intermolecular (as in protein binding)?

Folded proteins are typically compact, densely packed structures [144, 224]. The primary driving force of this compaction is the hydrophobic effect [88, 96, 435, 574]. The hydrophobic effect refers to the tendency of aliphatic chemical moieties to reduce their solvent accessible surface area. The thermodynamic driving force for this burial of hydrophobic amino acids (typically Ile, Leu, Val, Met, Try, Phe, Trp) originates from the entropic and enthalpic cost of solvating aliphatic groups. Consequently, the burial of hydrophobic groups should

not be thought of as a tendency for aliphatic groups to be attracted to one another by some ethereal 'sticky hydrophobic force', but instead reflect their repulsion from aqueous solutions. Importantly, if we take this to its logical conclusion - that protein folding is driven by the hydrophobic effect, which in turns leads to a hydrophobic core in the majority of folded proteins - then it should be clear that protein folding is not necessarily intrinsic to all polypeptides, but in fact reflects the interplay between chemical groups associated with the amino acid sidechains and the surrounding solution environment. The importance of solvent context will be explored in great depth in chapter 5.

Hydrophobic groups are not the only determinant of protein compaction. Amino acids with amide sidechains (Asn and Gln) can engage in hydrogen bonding with one another and themselves to drive chain compaction. Charged residues can engage in intramolecular salt bridges - strong, persistent electrostatic interactions that can lock two distal parts of a protein together [310]. However, charged residues will also show strongly repulsive or attractive interactions with one another in a conventional Colombic fashion, and are typically found on the exterior of proteins due to their extremely favourable free energy of solvation [72, 621]. Consequently, we speculate that charged residues may have a strong influence on protein topology, where the native state must balance both the externalization of charged residues with the efficient distribution of like-charged residues across the protein surface to minimize electrostatic repulsion. Aromatic residues may engage in pi-pi or cation-pi interactions, either between sidechains or even with the peptide backbone [118, 190, 191]. Beyond these pairwise interactions, many residues show distinct preferences for specific types of secondary structure. For example, proline and glycine are typically not found within helices, but often in flexible loops connecting distinct regions of secondary structure [17, 112].

Many of the factors here are also important in mediating intermolecular protein-protein interactions. Charged residues frequently form binding interfaces, where a combination of protein geometry and electrostatic interactions allow for high specificity and high affinity interaction sites [303,537]. Similarly, exposed hydrophobic moieties allow for the formation of obligate binding sites, often critical in the formation of complex protein assemblies [103,271]. For example, the leucine zipper motif is a string of hydrophobic residues along one face of an α-helix. This hydrophobic surface drives helix dimerization to shield those residues from the solvent [432]. Indeed, in much the same way that solvation forces are often considered critical for protein folding, protein-protein interactions can be predicted and understood to a high degree of fidelity using a similar solvation-centric framework [103].

There are many different factors that influence the folded structure of a protein and that determine 'foldability'. For globular proteins, a hydrophobic core is important, although not strictly necessary (see the associated discussion in chapter 2). An important idea is that the folded state is significantly stabilized by secondary structure, which provides a well-defined locally stable arrangement for the polypeptide backbone, allowing the sidechains to mediate the additional contacts that dictate the final tertiary structure.

1.4 Mechanisms of Protein Folding

In this final subsection, we provide a brief overview of some of the ideas and concepts in protein folding. It may seem odd to include a discussion on protein folding as part of a thesis that is (primarily) focussed on proteins that don't fold. The goal in doing so is to make it clear that *all* proteins - folded proteins, partially folded proteins, unfolded proteins - are subject to the same thermodynamic driving forces. The various mechanisms proposed/hypothesized

to explain protein folding are just as relevant for thinking about the kinds of interactions that occur in disordered proteins, and so in many ways the protein folding problem is of direct relevance to understanding intrinsically disordered proteins (see chapter 2).

Understanding the mechanisms through which proteins fold has remained an open and important question in protein biophysics for over fifty years [73, 74, 143, 154, 198, 280, 431]. Although the primary sequence clearly encodes the native state structure for many foldable proteins, simply encoding this information is insufficient to guarantee the actual formation of the native state in solution. This was elegantly demonstrated by Cyrus Levinthal in 1969, where he suggested a simple but profound thought experiment [326]. The number of possible unique configurations associated with a polypeptide chain of even modest length (say 200 amino acids) is $\sim 1^{100}$. Even after all the states that would introduce steric collisions are removed from the set of possible states, we are left with a problem: if a chain sampled each conformation sequentially in a non-redundant manner at the fastest possible rate of interconversion (say 1 ns per state) the time taken to fully explore conformational space to find the native structure would on average take longer than the age of the universe. Yet, proteins do spontaneously fold in microseconds with remarkable fidelity. Consequently, protein folding cannot simply be a random search, but instead requires a mechanism (or a collection of mechanisms) to aid in the search for the native state. Various putative mechanisms have been proposed to provide an atomistic description of the events that allow an unfolded polypeptide to rearrange itself into the native state.

An important feature of folded proteins with direct relevance to the folding mechanism is that they display cooperative stability. This is characteristically described by experimentally determining the fraction folded vs. a perturbant (e.g. concentration of denaturant or temperature) and observing a sigmoidal unfolding curve (see fig. 1.2). The presence of

cooperative conformational behaviour is not necessarily indicative of folding in the conventional sense, and that cooperative folding is not a requirement for the acquisition of protein structure [418, 524]. However, that the majority of proteins display cooperative folding, this is at least consistent with the interactions that stabilize the native state typically behaving in a hierarchical manner.

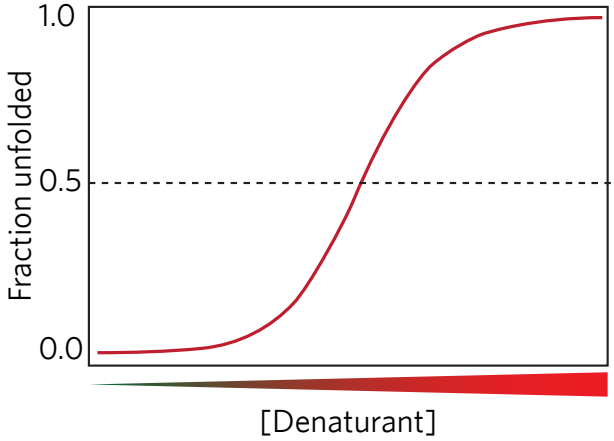

Figure 1.2: Schematic of a conventional protein folding/unfolding curve. The fraction of folded protein is assessed at various concentrations of denaturant. The sharp transition is consistent with the folded state behaving in a cooperative manner.

A second important idea is the folding pathway - a route through conformational space that allows a protein to transition from a denatured state to a folded state [73, 74]. Folding pathways are broadly consistent with many different types of folding mechanism. For multi-state folders, the transition states that define the rate-limiting bottlenecks associated with the folding process can be thought of as being directly related to locally metastable states along a folding pathway, commonly described as folding intermediates [25, 262, 491]. With the advent of Markov state models (MSMs) to describe protein folding, the states identified and the flux between different states provides a tangible realization of what those folding pathways may mean at atomistic resolution [61, 101].

The **sequential folding model** suggests that proteins fold via a well-defined series of steps. along a strictly linear reaction pathway [596]. Evidence for multiple folding pathways suggests there need not be a single route for folding, but the sequential model does provide a simplifying framework amenable to analytical interpretation [4, 62, 339]. One approach for describing the kinetics of protein folding is to use Kramer's theorem, which describes the folding process as a diffusion reaction on a 1D energy landscape with a single barrier corresponding to the transition state [105,305]. Such a description treats the folding pathway as a sequential two state reaction with a single transition-state species. However, this description is more relevant as an analytical model to describe the kinetics of folding than it is a means to obtain an atomistic description of the events that occur along the folding pathway, and apparent two state folding is entire consistent with multiple folding pathways [62].

The **nucleation-growth** model assumes there a local regions within a polypeptide where native structure is able to form *de novo*, and these nucleation sites serve as seeds for the formation of global structure [638]. Importantly, multiple nucleation sites can drive the formation of local structure independently of each another, allowing folding to proceed via a divide-and-conquer style mechanism, a result consistent with theoretical and experimental results for a number of proteins [214]. An important signature of the nucleation-growth model is a pre-nucleation phase (lag phase) and a post-nucleation phase (growth phase). The nucleation-growth model makes no assumptions about the nature of the nucleation site(s), making it an attractive model in part since it is consistent with one or more of local, long-range, secondary, or tertiary structure elements driving folding. However, the conventional interpretation of the nucleation-growth model does not allow for folding intermediates (long-timescale states in a nucleation-growth model originate solely from the lag phase - folding

proceeding exponentially during the growth phase)[4]. Given the abundance of known folding intermediates, the nucleation-growth model is clearly inadequate as a *universal* description of protein folding, although this does not proclude it being a good model for certain proteins [174].

The **diffusion-collision-adhesion** model, developed by Weaver and Karplus, shares several features with the nucleation-growth model [279, 280]. In the diffusion-collision-adhesion model, short local regions of the polypeptide chain condense and coalesce to form metastable structures. These micro-domains fold and unfold, but once formed they may collide with other meta-stable micro-domains to give rise to larger and more stable condensed local structure, driving a hierarchical folding process. Testing the validity of this model has historically involved altering the solution viscosity, although more recent work on internal friction suggests there are complicating factors that may convolute the interpretation of these results [554]. In the original formulation of the diffusion-collision-adhesion model, micro-domains were considered to be local with respect to the polypeptide chain. However, various results have shown that long-range interactions are present even in the unfolded state, suggesting that micro-domains may include clusters of residues that are far away from one another in sequence space. We speculate that the diffusion-collision-adhesion model could be recast as a phenomenon akin to phase separation near the critical point with the constraint of chain connectivity, where the kinetics might be expected to behave in a manner analogous to a (bias) diffusion-limited Ostwald ripening [332, 619].

The **framework model** considers the folded state to be entirely determined by secondary structure elements, and suggests different secondary structure components can assemble

[4]A second interpretation of the nucleation-growth model states that the once formed, the nucleus represents a metastable intermediate that is observable before complete folding occurs. This description is incompatible with convention nucleation-growth kinetics for crystals or polymers, so it is easier to consider this version a separate mechanism entirely [78].

independently and then 'lock together' [465, 466]. In this model, the tertiary structure is an emergent property of interactions that occur once local topology has been determined by secondary structure. This model can be considered a limiting case of the diffusion-collision-adhesion model, in which micro-domains are secondary structure elements. As far as we can tell, there is no intent to suggest that that the secondary structure elements must remain physically separate from one another during their initial formation. However, this model requires that the majority of secondary structure elements form before tertiary structure, a result inconsistent with various results that show secondary structure formation can occur at a range of points along the folding pathway [174,339]. However, the original 1973 framework model paper by Ptitsyn was the first to suggest that partially stable intermediates may form, a prediction shown to be true for many (though not necessarily all) proteins [166, 288, 465, 480].

The **hydrophobic collapse** model suggests that a protein will undergo rapid compaction around the hydrophobic residues to form a dense globule [141]. This globule will then re-arrange to form the native state. Such a model, contrary to the framework model, suggests that secondary structure forms as an emergent property of the collapsed state, which in turn is determined by the distribution of hydrophobic residues [142]. Elegant work by Lin & Zewali demonstrated that the reduction in conformational space provided by hydrophobic collapse alone is sufficient to make folding a tractable exercise from a search perspective, but that this effect holds only up until domains of around 200 residues or so, consistent with the typical limit in size for single domain [334]. A conceptual challenge with the original hydrophobic collapse model is that the diffusion of a polypeptide chain within a collapsed globule is expected to experience substantial internal friction, such that backbone and sidechain reorganization may be substantially hindered and occur on timescales slower than those necessary for fast protein folding [554, 555]. To account for this, an updated

version suggests rapid compaction occurs concomitant with secondary structure formation leading to the formation of a molten globule [599]. Regardless, such a model suggests substantial compaction precedes protein folding, a result consistent with some FRET studies but inconsistent with many SAXS studies [59, 255, 661, 673].

The **nucleation-collapse** model unites features from the nucleation-growth and hydrophobic collapse models [120, 214, 295]. The nucleation-collapse model suggests that local and long-range hydrophobic interactions form in the transitions state to help stabilize secondary structure. The transition state is also expected to contain native tertiary structure, and appearance of the native contacts associated with tertiary structure can be probed by ϕ-value analysis [121]. Hydrophobic interactions are responsible for driving folding, but local and specific native-like micro-domains form early in the folding process and can form identifiable intermediates. This nucleation-collapse mechanism is relatively consistent with work from Englander describing local regions within proteins that continually fold and unfold even in the native state, referred to as **foldons** [166, 353]. Foldons represent discrete structural units that experience both long and short range interactions. A key difference between a foldon-centric model and the nucleation-condensation model is that foldons fold in a specific and stepwise pattern, while the nucleation-collapse model makes no assumptions regarding the order of nuclei formation. The emergence of a specific sequence of folding nuclei is not inconsistent with the nucleation-collapse model, but rather the nucleation-collapse model could be considered a highly permissive folding model of which a foldon-centric model is a specific subclass. The identification of foldons via hydrogen exchange (HX)-based methods has fairly rigid thermodynamic and kinetic requirements, which in itself may suggest they exist only for a certain subclass of folding pathways where these requirements are met [353].

Beyond specific mechanisms, an idea that has become pervasive throughout the protein folding literature is the concept of the 'folding funnel' [143]. The folding funnel refers to the free energy landscape explored by the protein, and by design implies that the native state(s) represents a single global minimum with relatively smooth edges (see fig. 1.3). As a consequence, a wide variety of unfolded states will converge towards a single native basin, which may represent a collection of folded states or a single native conformation. Folding intermediates would be described by local minima along the funnels sides, and represent transient meta-stable states. The stability of the intermediate depends on the depth of the associated local minima. The principle of mimimum frustration provides an explanation for the smooth sides of the funnel, and predicts that evolution selects for amino acid sequences that improve the smoothness of the funnel [73, 74, 173, 648]. While an attractive hypothesis, folding fitness is far from the only selection determinant in protein evolution. It remains to be seen if true fitness[5] is substantially impacted by the 'ruffling' of this folding funnel, where ruffling refers to the introduction of more locally stable intermediates.

In the last fifteen years or so, all-atom simulations have provided unparalleled resolution for directly examining putative folding mechanisms [61, 62, 185, 186, 339, 453, 617]. While these simulations represent finite descriptions of a simplified model of our understanding of some of the relevant physics, they also allow us to ask specific and well defined questions in a controlled manner. The development of specific hardware and distributed-computing software to allow the construction of massive trajectories (either via single long simulations or by MSM reconstruction) has allowed direct observation of folding reactions for many small single domain proteins [61,535]. Work by Bowman and colleagues examining NTL9$^{1\text{-}39}$ and the Villin headpiece (HP-35) demonstrated that for these small single domain proteins,

[5]True fitness refers to the idea that true selective pressure is an emergent property of an organism's dynamical environment and is likely poorly replicated by any kind of *in vitro* assay due to the fact that fitness represents stability to uncertainty, and scientific assays are, by definition, controlled.

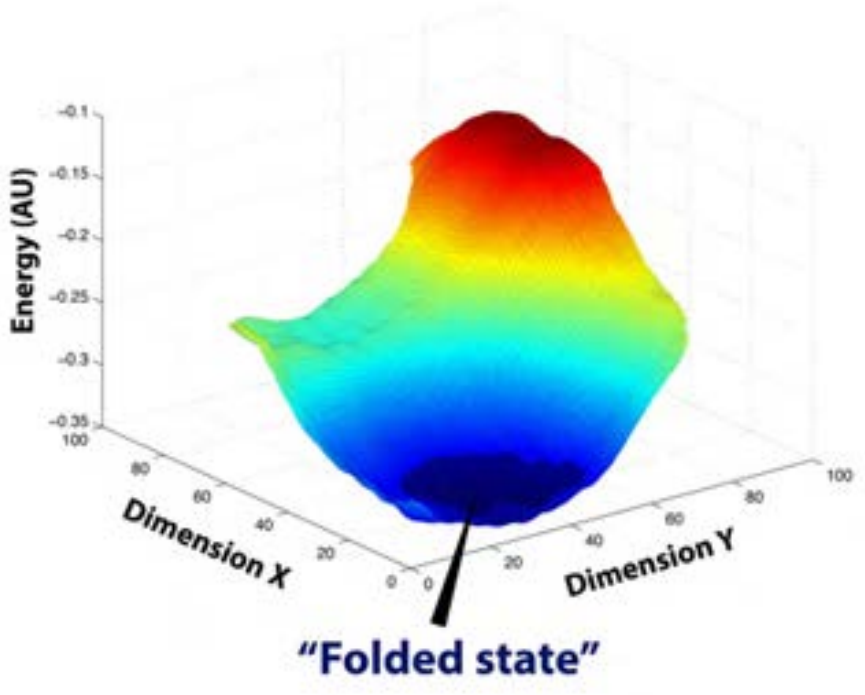

Figure 1.3: Schematic of a hypothetical folding surface for a typical single-domain protein. The folding funnel represents the energy landscape associated with a protein's conformational space. The funnel has a single global minimum that defines the native state, and relatively smooth sides that ensure that the global minimum is always reached.

the native state behaves as a kinetic hub, where interconversion between non-native and native states occurs more frequently than between different non-native states, and suggests that multiple different pathways can lead to the formation of the native state [62, 617]. Work by Best and colleagues analysing several folding trajectories generated by Shaw *et al.* strongly suggests that in these examples non-native contacts play almost no role in the folding transition state, although it is reasonable to ask if this is perhaps a defining feature of small, fast-folding single-domain proteins [45]. A potential pitfall of these trajectories is that the unfolded states have approximately the same global dimensions as the folded state, a result that is largely inconsistent with an extensive body of evidence examining the unfolded state

of proteins (see chapters 5 and 7). In more recent studies examining diffusion through the folding transition state, there is evidence that native and non-native electrostatic interactions act to slow the folding process, offering direct insight into the atomistic determinants of folding kinetics [105]. In summary, from watching a sixteen residue β-hairpin peptide fold and unfold to simulations of million-atom macromolecular complexes, all-atom simulations have fundamentally changed how protein folding questions are asked [438, 448]. Simultaneously, equivalent advances in protein structure determination and florescence based approaches have re-shaped the experimental landscape in terms of the accessible temporal and spatial resolution [24, 521].

Recent advances in microfluidic mixers, photo-detectors, computer hardware, and computational models are allowing us to enter a new phase in the study of protein folding. The conformational steps associated with the folding process can be followed directly and in real time both experimentally and computationally. The diversity of results suggests that there may not be a single protein folding mechanism. In much the same way as proteins fold into a wide range of different structures, protein folding may proceed via a range of different mechanisms. This is already evident in the fact that many larger proteins require chaperones to aid in folding, suggesting that due to a range of possible factors (cellular milieu, protein size, protein topology, protein sequence *etc.*) not all proteins are able to reach their native state autonomously [290, 511]. A final point not explored in the preceding sections is the relationship between *de novo* folding and co-translational folding. Protein synthesis occurs substantially slower than the majority of folding processes observed to date [240]. How does the ribosome influence folding? This remains an open question, with hypotheses from topology dictation to the modulation of the nascent chain's solubility [258, 419, 580]. We introduce these ideas only to make the point that we are far from 'done', in terms of understanding the protein folding problem.

1.5 Summary

We have provided a brief overview of the molecular architecture associated with polypeptides, the structural features associated with folded proteins, and a short overview of protein folding. While this thesis does not directly explore the various folding pathways described here, understanding them provides some useful context for the work in chapters 7 and 5. It also helps illustrate that all proteins (folded or otherwise) are heteropolymers made up of the same types of amino acids and subject to the same types of chemical interactions. With this in mind, there is no reason why the sequence determinants that influence the conformational behaviour of intrinsically disordered proteins should be any different than those that influence the conformational behaviour of unfolded or partially folded foldable proteins. Recent work examining the dynamics during folding found non-specific electrostatic interactions were a major determinant of intramolecular interactions, entirely analogous to the behaviour observed in disordered proteins [105, 359, 364, 405]. Similarly, in unpublished work we have found that the global dimensions of the unfolded state of a foldable protein directly correlate with the net charge per residue, mirroring behaviour in intrinsically disordered proteins[6] [359, 364, 405]. In conclusion, these results suggest a general framework for relating amino acid sequence to conformational behaviour should be broadly applicable to all proteins, regardless of if they are at equilibrium or not.

[6]Peran, Holehouse, *et al.* (unpublished)

Chapter 2

Intrinsically Disordered Proteins

2.1 Introduction

As discussed in chapter 1, many proteins fold into well-defined three-dimensional structures, where that structure is critical for normal cellular function. This is true for many proteins, but for many others it does not represent the complete story.

Given the structure-function paradigm - in which the three-dimensional structure of a protein determines function - a tempting leap of scientific faith is to extrapolate that structure is required for function. Indeed, this perspective was broadly held by biophysicists and structural biologists through much of the 1970s and 80s, and with good reason; the advent of crystallography had provided a powerful window into the atomistic world of molecular function, and solved mystery after mystery. In May of 1988, Paul Sigler - a card carrying structural biologist - published a short News & Views piece in *Nature* on the topic of transcriptional activation domains (TADs). In the opening section, Sigler made the following, somewhat heretical, but extremely prescient statement [543]:

> *The more that is known about the amino-acid sequences of proteins that participate in transcriptional activation, the clearer it becomes that many of the critical events cannot depend on the precise complementarity that we associate with the interactions of globular proteins during molecular assembly, and the biding of substrates, cofactors, and haptens. The latest entries in this chronicle of molecular impression are the activator domains of the proteins that stimulate transcription by RNA polymerase II. These activating 'structures' (and one must use that term advisedly) are targeted to specific DNA-sequences usually by a specific DNA-binding domain on the same polypeptide. The DNA-binding domain appears to have well-defined structural motifs; by contrast, mutational studies on the activator domains suggest a disquieting picture of a conformationally ill-defined polypeptide that can function almost irrespective of sequence, provided only that there is a sufficient excess of acidic residues clustered or peppered about.*

At the time of writing, this was a somewhat controversial view. It was well established that many proteins (or regions within proteins) failed to fold *in vitro*, to the chagrin of X-ray crystallographers. A common approach was to identify regions within a sequence that appeared devoid of hydrophobic residues - the residues known to drive folding through the hydrophobic effect - and simply excise these 'unfoldable' regions from the sequence. Conveniently these regions were often identified in the N- or C-termini, so a shifted start codon or premature stop codon was often enough to convert a poorly behaved protein into a truncated yet soluble species that would be far more amenable to crystallization and further structural characterization. Moreover, when the amino acid sequences associated with these regions were compared across different species the degree of conservation was found to be much poorer than in the well behaved regions, suggesting a lack of evolutionary selection pressure. The (entirely reasonable) conclusion was that these regions were dealt with by the cellular proteostatic machinery, such that they folded in the context of the cellular environment, but absent those factors remained unstructured, driving aggregation and poisoning crystallographic screens due to their inability to form regular structures.

Sigler hypothesized something different. His suggestion that these regions are functionally relevant yet exist as a "conformationally ill-defined polypeptide" seemed paradoxical in a world where function required structure. However, work over the next decade helped to uncover that although function is determined by structure, structure need not mean folded. In fact, the 'failure' to adopt a folded conformation in no way precludes critical function.

How should we refer to these 'conformationally ill-defined polypeptides'? The term 'unfolded' is too deeply associated with the protein folding literature, where unfolded reflects the state under high concentrations of denaturant. Unstructured, though often used, introduces an unfortunate implicit binary classification (structured vs. unstructured) which is at best inaccurate and at worst misleading. The field has largely settled on the term 'intrinsically disordered', giving rise to the nouns intrinsically disordered proteins (IDPs) and intrinsically disordered regions (IDRs). For the remainder of this thesis we will collectively refer to these proteins and protein regions as IDPs. Despite the prevalent use of the term IDPs, there are relatively few proteins that are entirely disordered; in most cases proteins consist of folded domains coupled to disordered regions, although the numbers and relative sizes of the various subregions (folded and disordered) vary widely. IDPs range from 10-15 residues to thousands of residues, and participate in a diverse set of cellular functions [21, 158, 350, 542, 584–588, 605, 608, 654].

What does disorder mean to a protein? In expanded homopolymers such as dextran or PEG, where there is no expectation of structure in terms of a fixed conformation, disorder refers to random-coil behaviour, where the ensemble averaged behaviour on both local and global scales is well described by polymer theories. Proteins are not simple homopolymers; different combinations of amino acids impart a rich repertoire of functional chemistry leading to exotic emergent properties. Consequently, while ensemble average properties that capture global

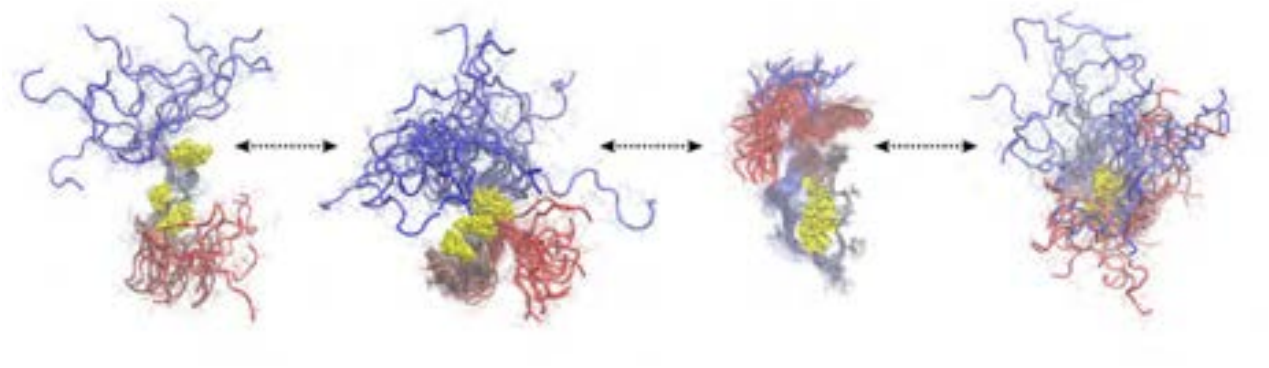

Figure 2.1: IDPs exist as an ensemble of conformational states. Shown here is an example of sub-states extracted from a simulation of the GCN4 transcriptional activation domain. Each of the four states represents a sub-ensemble of states extract at random from a much longer simulation. These are not 'representative' structures - if four additional states had been selected at random four new states would be found - but they do provide some sense of the heterogeneity associated with an IDP's conformational ensemble

protein behaviour (average dimensions, average end-to-end distance, *etc.*) can be described by simple polymer physics models, this does not necessarily mean that those models are meaningful or predictive [189, 507].

A crucial discovery over the last ten years (and a major focus of this work) is that despite being disordered, IDPs typically display complex and well defined behaviours including the formation and loss of transient secondary structure, local compaction or expansion, and local and long-range interactions. These seemingly 'unstructured' proteins are, in fact, full of 'structure', in an information-theory sense. The amino acid sequence of an IDP directly encodes the intrinsic properties of that protein's conformational ensemble. In turn, those properties can be dramatically altered by a variety of factors including binding partners, solvent conditions, and post-translational modifications.

The remainder of this chapter is divided into several section. First, we will review the relationship between amino acid sequence and conformational behaviour, considering the sequence determinants of disorder. Next, we will outline a selection of experimental approaches used in the following chapters to studying IDPs. Finally, we consider theoretical and computational approaches for characterising IDPs, notably simulation approaches and analytical models.

2.2 Sequence Determinants of Conformational Behaviour

The study of folded proteins has been driven by structural data. We can generate mechanistic hypotheses, explain the effect of mutations, and construct models to explain complex phenotype rooted in precise, quantitative information [2, 5, 629]. For IDPs, using a single protein 'structure' to gain insight into mechanism is not an option. The absence of high resolution structural information is not a weakness of our current methods, but a fundamental property of the system itself. IDPs exist as an ensemble of states; it is temping to use clustering methods to identify the most commonly populated conformations and treat these as structural hubs, but this enforces an artificial and semi-arbitrary discretization of the ensemble. A better option is to use to a statistical language to describe these ensembles in terms of the distribution of values associated with some order parameter of interest.

As discussed in chapter 1, the thermodynamic driving force behind protein folding is generally considered to be the hydrophobic effect. The sequestration of hydrophobic residues (notably Ile, Leu, and Val) into protein interiors drives compaction, and the formation of regular secondary and tertiary structure gives rise to an energetic minimum associated with the folded state [143, 371]. IDPs, on the other hand, are typically depleted in these bulky

hydrophobic residues, but enriched in polar, charged, and proline residues [603,608]. Without a strong driving force to populate a single well-defined global minimum, IDPs instead explore a heterogeneous ensemble of different states.

The absence of hydrophobic residues and enrichment in charged/polar residues provides a sequence-based rational for the inability of IDPs to fold. A consequence of these robust sequence features is that there are many algorithms that can provide reasonably robust sequence-based predictions of a protein/regions degree of disorder. These algorithms allow local regions of disorder to be predicted from sequence alone, and several meta-prediction servers (servers that combine results from multiple different predictors) such as D2P2 (http://d2p2.pro) and MobiDB (http://mobidb.bio.unipd.it/) have emerged as critical tools for working with protein sequences [148, 149, 426, 461].

An absence of hydrophobic residues does not *necessarily* preclude folding. Examples include the snow-flea antifreeze protein (sfAFP) which we will return to in chapter 5 and the curious *S. aureus* cell wall adhesion protein SasG (see fig. 2.2) [193,208]. Both proteins are predicted to be disordered with high confidence (in the case of SasG, the region of interest consists of around 40% charged residues, where folded proteins are typically around 15% charged). However, both have been crystallized and shown to form well defined folded structures that simply lack a hydrophobic core. While these are likely the exception and not the rule, they provide a useful reminder that our ability to predict disorder is far from perfect. They also highlight the fact that a collapsed hydrophobic core may be a convenient mechanism for proteins to fold and to evolve, but is not the only way in which a well defined three-dimensional fold can be achieved [50].

If folded proteins are typically compact due to their hydrophobic core, we can already see that there may be a relationship between amino acid sequence and conformational behaviour. The

Figure 2.2: SasG and sfAFP are unusual folded proteins, in that they have amino acid compositions consistent with a disordered protein but have been shown to form a well defined folded structures.

holy grail of structural biology is to achieve robust and rapid *de novo* full protein structure predictions from sequence alone. With the ROSETTA ecosystem this is approaching a reality, although a significant investment of time is still required [64, 242, 433]. The equivalent goal with IDPs is to generate accurate predictions of an IDPs conformation ensemble based on amino acid sequence alone.

We can re-state this goal as a well defined question: how does sequence determine conformational behaviour? Let us consider this question across two length-scales: global conformational behaviour and local conformational behaviour.

2.2.1 Global Conformational Behaviour of IDPs

For simplicity, let us begin by thinking about global properties, such as the ensemble average radius of gyration (i.e. ensemble size) and the ensemble average asphericity (i.e. ensemble shape). Both these parameters are formally defined in chapter 5, and provide an a general expected-value description of the size and shape of an IDP. If all IDPs behaved according to standard homopolymer models, such as the Gaussian chain or self avoiding walk (SAW) then the asphericity and the radius of gyration would depend only on the number of amino acids in a sequence. In reality, these global properties show well-defined sequence dependencies, suggesting that different amino acid compositions engender different conformational behaviours.

IDPs enriched in certain polar amino acids (notably glutamine and based on aggregation behaviour likely asparagine), aliphatic amino acids (alanine, methionine, leucine, isoleucine, valine) and aromatic amino acids (phenylalanine, tyrosine and tryptophan) tend to be more compact, to the extent that they can behave as disordered globules [116, 364, 404, 421, 483]. This does not mean they fold - these compact conformations undergo exchange between different compact states on some timescale (see fig. 2.3), but they are a far cry from random coils thrashing about in solution.

A prime example of a polar rich IDP is polyglutamine, which has been shown through experiment and simulation to form compact globules [116, 117, 277, 627, 644]. Similarly, the N-terminal domain of Sup35 forms compact globules mediated at least in part by interactions between glutamine and asaparagine in the N-terminal prion domain [404]. Although hydrophobic residues are less common in IDPs, the P-domain of poly(A)-binding protein (Pab1) is highly enriched for several different hydrophobic residues and undergoes robust

collapse into a disordered globule [483]. Furthermore, sequence permutants expose a direct relationship between hydrophobicity and the radius of gyration, as measured by small angle X-ray scattering. The impact of other polar residues is less clear. Polyglycine forms dense disordered globules (see chapter 5), but sequences enriched in glycine can be highly expanded (see chapter 12) suggesting a context dependence for the role of glycine [193]. This is discussed extensively in chapter 5.4.1. Serine and threonine are also less well characterized. There is some evidence that serine may promote chain expansion due to its inability to strongly interact with partners. Histadine is naively expected to interact strongly with amides (backbone and sidechains) via its nitrogen and also through its partial pi-system, but its ability to undergo protonation and become charged may allow it to reduce chain compaction.

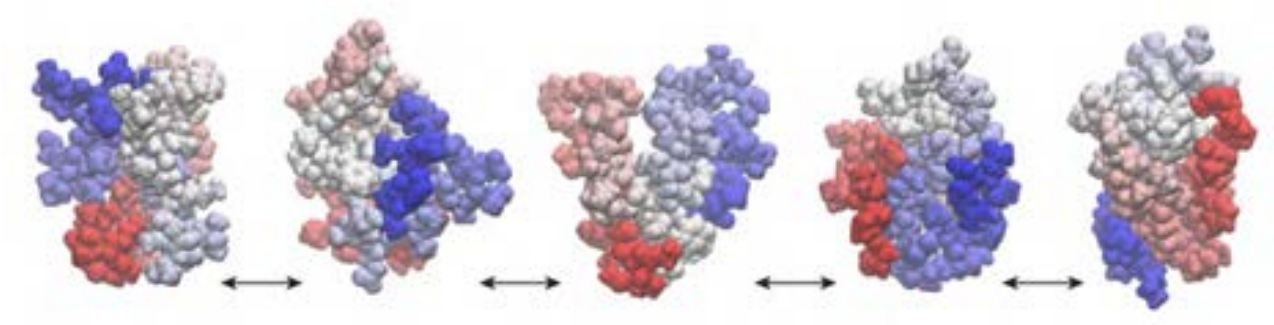

Figure 2.3: Examples of different compact globular conformations for a polar and hydrophobic rich IDP. Coloring runs from blue-to-red (C to N terminus)

IDPs enriched in charged amino acids are generally more expanded [359, 364, 405]. The origin of this expansion is two-fold. *Firstly*, charged sidechains have a highly favourable free energy of solvation. They are typically found on the surface of folded proteins, and in IDPs these favourable free energies of solvation make burial energetically expensive [19, 359, 405]. *Secondly*, like-charged amino acids experience electrostatic repulsions with respect to one another, and this repulsion can lead to chain expansion. For polyelectrolytic IDPs (IDPs

that are strongly enriched in either positively or negatively charged residues) understanding charge repulsion as a driving force for chain expansion is self-explanatory, and generally shows good agreement with theoretical descriptions from the polyelectrolyte literature [405]. However, in neutral but charged IDPs (polyampholytes) it is less clear if we should expect charge repulsion to be relevant, or if attractive interactions between oppositely charged residues will dominate, driving chain compaction. To explore this question, the impact of charge patterning - the distribution of oppositely charged residues - in IDPs was examined in a series of disordered peptides of a fixed amino-acid composition [126]. The patterning of charged residues can be quantified by the normalized parameter κ, which reports on the extent of mixing between oppositely charged residues. When $\kappa \approx 0$ oppositely charged residues are evenly distributed with respect to one another. When $\kappa \approx 1$ oppositely charged residues are fully segregated. For many polyampholytic sequences, charge patterning is an important determinant of global and local conformational behaviour. It provides a mechanism though which Nature can define and regulates cellular interactions, offering a high fidelity yet somewhat sequence independent mechanism to mediate inter and intra molecular interactions [125, 335, 336, 420, 436, 517]. For an extensive discussion on the parameter κ please see section 4.3.5.

Taken together, these insights begin to provide us with a first-order picture of how the amino acid sequence of IDPs influences their conformational behaviour. We can codify these insights into a diagram of states, as developed by Das and Pappu and shown in fig. 2.4a [126]. The diagram of states classifies a given IDP sequence based on the fraction of positively charged residues (f_+) on the x-axis and fraction of negatively charged residues (f_-) on the y-axis. Based on the charge composition the possible space of sequences is divided up into five distinct regions (R1-R5).

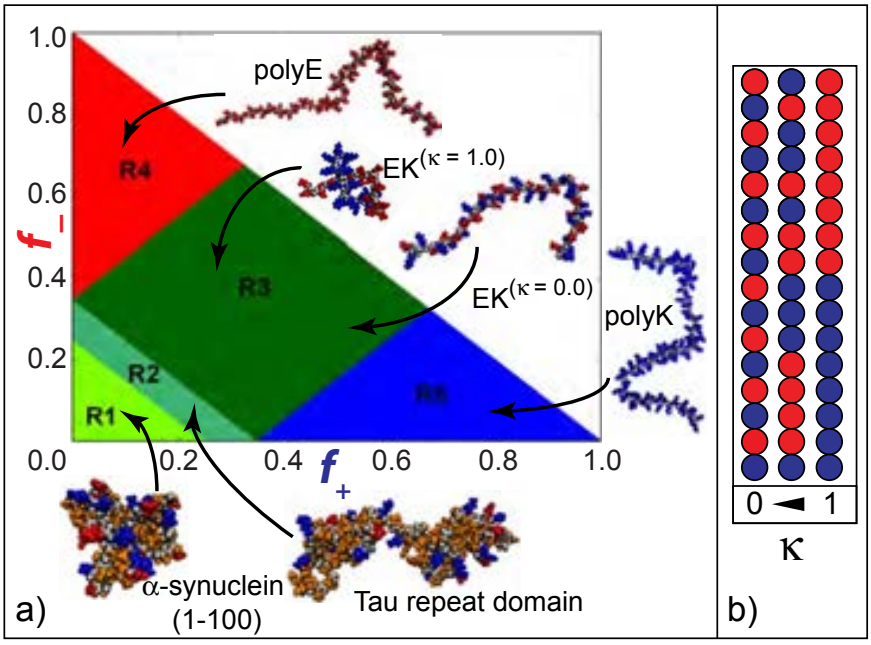

Figure 2.4: The diagram of states provides a coarse-grained classification tool for predicting conformational behaviour. Although originally suggested that sequences that fall into R1 would form collapsed globules, further work suggests that the determinants of collapse in disordered proteins is more complex than solely based on charge, and the prediction that a sequence that falls into R1 forms a disordered globule is likely misleading in many cases.

Polyelectrolytic sequences fall in R4 and R5 and are expected to be highly expanded due to charge repulsion and favourable sidechain solvation. Strong polyampholytic sequences fall into R3, and weak polyampholytic sequences into R2. The conformational behaviour of sequences in R3 is expected to be significantly determined by charge patterning, while those in R2 will likely be determined by a combination of charge interactions, polar interactions, and intrinsic secondary structure propensities. Sequences in R1 are generally devoid of charge residues, so are expected to form more compact ensembles, although this need necessarily mean the formation of disordered globules.

The classifications done by the diagram of states should be treated more as a convenient tool to generate an initial assessment of a sequence than as a rigorous method to predict global dimensions. The boundary lines should be considered distinctly fuzzy, and other factors (patterning of polar residues, presence and distribution of proline residues, glycine content) also appear to be highly relevant. As an example, there is mounting evidence that sequences with a low fraction of charged residues can still be relatively expanded [189, 366]. An additional challenge is raised by the fact that across an IDP of (say) 100 residues, one sub-region may fall into R1 (implying local compaction) while another may exist in R3 or R4, giving rise to sequence average properties that appear to place the sequence in R2, despite the fact that on some local level no part of the sequence is squarely in R2. This raises an obvious question: on what length-scale (if any?) should we subdivide an IDP sequence into to extract more meaningful information? This is an open question that is being pursued.

2.2.2 Local Conformational Behaviour of IDPs

The preceding section considered the relationship between amino acid sequence and *global* conformational behaviour, such as the radius of gyration or asphericity. These global order parameters represent a mean-field description of a polypeptides global behaviour taken over all length-scales. We must also consider *local* conformational behaviour, such as specific inter-residue distances or local secondary structure. These are conformational preferences that are only emergent in a heteropolymeric sequence, and arise from the chemical heterogeneity introduced by the amino acid sidechains. Many IDPs experience well-defined secondary structure preferences (typically helicity), which may be critical for function [57, 114, 221, 534, 653]. Although helicity is transient, it is often associated with a more ordered state

that undergoes interaction with a binding partner, frequently via a coupled-folding-and-binding reaction [534, 564, 653]. A second feature observed in IDPs is well defined attractive or repulsive interactions experienced across various different length-scales. While IDPs are (by definition) not folded, certain regions, motifs, or residues may experience anisotropic interactions (attractive or repulsive) leading to the emergence of well defined but relatively broad conformational preferences [513].

Are local and global conformational behaviour inherently coupled? While they certainly could be, this need not necessarily be the case, as illustrated by our work on Ash1. As is discussed in chapter 6, a combination of SAXS, NMR and simulations allowed us to examine changes to local and global conformational preferences induced by phosphorylation. In this system, despite well defined changes to the local conformational behaviour upon phosphorylation, the global conformational behaviour remained unchanged. This decoupling of global and local behaviour is reminiscent of the decoupled between end-to-end distance and radius of gyration, as observed for a series a IDPs [189].

Does local conformational behaviour influence function? In the case of the transactivation domain (TAD) of GCN4, extensive simulations coupled with a high-throughput method to assess the transcriptional activation driven by different TAD variants revealed that local conformational preferences have a direct impact on function[7]. In a separate study, we hypothesize that pH mediated changes in conformational behaviour are critical for the function of Sup35[8].

The preceding section raises obvious questions regarding the sequence lengths over which global properties are useful. More generally, can we divide IDPs into subdomains with

[7]Staller, Holehouse *et al.*, (submitted)
[8]Franzmann. ..., Holehouse *et al.*, (under review)

distinct properties? In folded proteins, subdomains are easily defined as regions that are structurally distinct from one another. Although these domains may engage in inter-domain interactions leading to complex macromolecular assemblies, subdomains can typically be excised from the native sequence and studied in isolation. An equivalent definition for subdomains in IDPs is challenging - if the disordered region engages in constant but transient interactions across the entire length of the IDP, then subdivisions may not be meaningful even if they are possible. However, for some IDPs convenient boundaries between regions with distinct sequence properties allows for an apparently clear delineation. One such example is the *S. cerevisiae* protein Sup35, which consists of a disordered and polar rich N domain, a disordered but highly charged M domain, and a folded C domain. While the N domain is compact, consistent with an absence of charged residues and enrichment of glutamine and asparagine, the M domain appears much more expanded [404]. In recent work not discussed in this thesis, we dissected the functional roles of the M, and N and C domains. Each subdomain imparts a distinct functional effect on the full length protein. The sequence architecture Sup35 from *S. cerevisiae* is convenient, in that the M and N domains have entirely distinct sequence properties. For *S. pombe* the delineation of regions between the N and M domain is much less obvious. Despite this, we identified specific sequence features that are conserved between the two fungi and we believe impart the same function in both cases. This at least hints that functionally relevant sequence features and *specific* amino acid sequence may be partially uncoupled in IDPs, and raises a possible route for sequence comparison of IDPs.

A final sequence feature deeply associated with IDPs that is not a focus of this thesis are Short Linear Motifs (SLiMs), also referred to as Eukaryotic Linear Motifs (ELMs) [130, 586, 609]. SLiMs are typically involved in mediating protein-protein interactions, and consist of 5-12 residue regions that are recognized by a cognate binding partner. These motifs are frequently

associated with sites of post-translational modifications, where their activity may be influenced by the status of those modification sites. The prediction of SLiMs remains a major challenge: even within SLiM databases there are examples of motifs which in some context are *bona fide* SLiMs but in others behave as true negatives. We will not focus on SLiMs for the remainder of this work, in part because they are typically associated with functional regulation within the cellular context, which is not a direct focus here. Nevertheless, they should not be ignored, and will have important implications in the determinants of function and fitness.

2.2.3 Evolution in IDPs

A common metric for evaluating the functional importance of some protein of interest is to perform an evolutionary comparison between a set of orthologous proteins taken from distantly related species [416]. The logic behind such an approach is reasonable; if orthologous proteins are similar in sequence then there must be have been a strong evolutionary pressure to maintain that sequence across many millions of years of divergence, implying a critical link between sequence, function, and fitness. The implicit inverse assumption from this is that for sequence that shown substantial divergence there is limited evolutionary pressure, suggesting these regions may not be important.

As highlighted earlier in this chapter, implicit assumptions can be dangerous. IDPs generally (although not always, see fig. 2.5) show relatively poor sequence conservation. If we consider this result in the context of the framework outlined above, this implies that these regions are under weak selective pressure and are unimportant for function. We *know* this is not

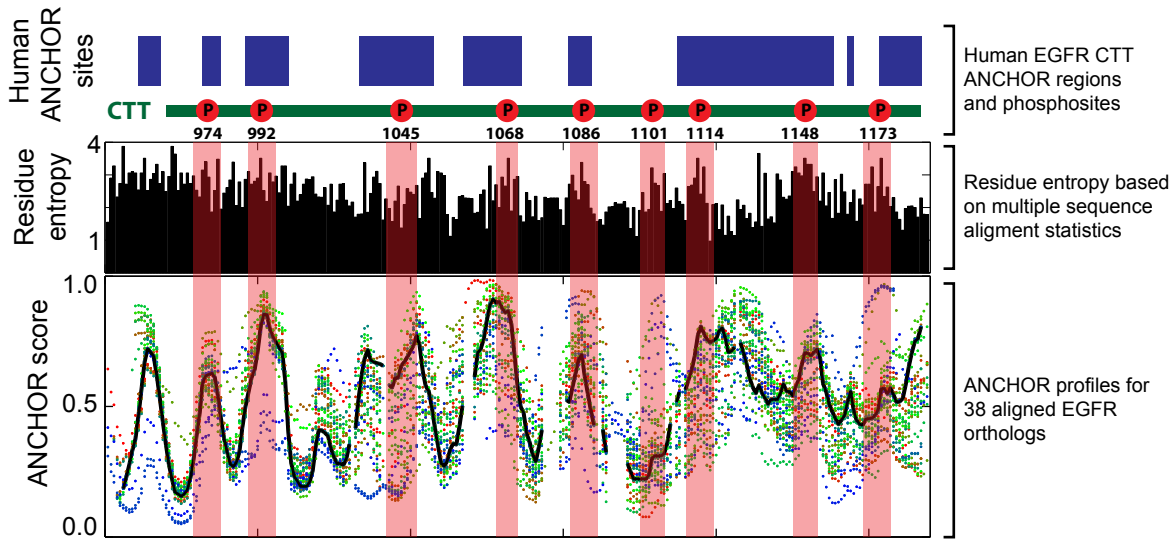

Figure 2.5: Although IDPs are typically fairly poorly conserved, the disordered C-terminal tail (CTT) of the epidermal growth factor receptor (EGFR) shows surprisingly good conservation across many different organisms. The tail acts as a signalling regulation hub, with different downstream effectors binding to different phosphosites along the sequence. Sequence analysis with the ANCHOR algorithm shows a strong coincidence of phosphosites and putative ANCHOR sites, with several other regions also displaying conserved and high ANCHOR scores suggesting at additional protein-protein interaction sites along the sequence. One possible interpretation of this is that strong sequence conservation in IDPs reflects a necessary co-evolutionary constraint imposed by a folded binding partner.

the case - disordered regions are frequently mutated in disease, and are required for normal function [605, 654]. How can we reconcile these two results?

In folded proteins there is an intrinsic and relatively tight coupling between amino acid sequence and cellular function, and hence fitness. This coupling is mediated by the folded structure. A small number of single point mutations can entirely ablate function, suggesting

folded proteins represent a relatively deep well in evolutionary space - small perturbations to the sequence can lead to a precipitous drop in fitness through loss of structure. In contrast, the lack of a well defined structure associated with IDPs means that the evolutionary minima they sit in is much more shallow. Single point mutations *may* disrupt function, but if they do not significantly alter the local sequence composition these mutations may be entirely benign. Consequently, the looser coupling between sequence and function associated with IDPs leads to a much broader sequence space of equivalent function, allowing large-scale sequence changes to have a minimal impact on function. In this way, IDPs can drift through sequence space as a function of time, while still existing under exactly the same tight selection as their associated folded domains and performing the same critical functions. Recent evolutionary analysis by Riback *et al.* provides an elegant demonstration of this; a disordered region in the highly conserved protein Pab1 shows poor sequence conservation, yet the composition is highly conserved [483]. The implications of this will be discussed in chapter 10

A final note that we will not dwell on further is the fact that IDPs are typically not found in prokaryotes. While around 30% of eukaryotic proteomes are disordered, this number hovers around the 2-5% mark for prokaryotes [437, 630]. It is unclear why prokaryotes seem so unwilling to take advantage of disorder, but may reflect genomic size constraints, hypervariable growth conditions, or less complex regulatory networks.

2.2.4 Function of IDPs

Much of this work is focused on the relationship between amino acid sequence and conformational behaviour. However, it is important to bear in mind that IDPs are not exploring interesting conformational behaviour for arbitrary reasons. Instead, those conformational

behaviours are often linked to their role in a smörgåsbord of biological functions. IDPs are frequently associated with cellular signalling, in part because they provide a convenient scaffolds for interaction sites [654]. These sites may encompass SLiMS that facilitate specific and/or regulated interactions, or may contain more promiscuous sites involved in the assembly of larger multi-component complexes. The structural plasticity associated with IDPs also allows the same binding motif to interact with multiple structurally distinct cognate partners [606]. Structural plasticity is frequently coupled with chemical plasticity in the form of post-translational modifications, which are enriched in disordered regions [252]. Taken together, eukaryotic signalling systems are able to take advantage of complex and highly regulatable assemblies that can facilitate adaptation by integrating multiple signalling pathways, with disorder playing a key structural and functional role in this process. In chapter 6 we will study a disordered region of the protein Ash1, a key transcription factor and signalling molecule involved in mating type switching in *S. cerevisiae*.

Disordered regions are often involved in mediating higher order assemblies. As will be discussed extensively in chapter 3, the formation of membrane-less organelles is often (though not necessarily) associated with disordered regions [27]. Signalling platforms - large assemblies of signalling proteins - are frequently believed to be driven by the presence of disordered regions. The assembly of the eukaryotic transcriptional initiation complex is associated with RNA POL II clusters of around 80 molecules, behaviour that is believed to be driven by the disordered C-terminal tail (CTD) [100,107]. In chapter 11 we will examine the C-terminal domain of the transmembrane protein Nephrin, and discuss how its partner-dependent phase separation provides a model for the dynamic assembly of signalling clusters on the membrane. In unpublished work, we identified a solution-responsive low complexity domain that drives robust protein assembly in an entirely tunable manner, suggesting that the grammar

of IDP-mediated phase separation is both rational and interpretable [9]. For phase separation and gelation that is mediated by folded domains connected by flexible linkers we found that the intrinsic properties of those linkers can fundamentally change the phase behaviour, highlighting a modulatory role for assembly in signalling complexes [223].

IDPs can also function in a more structural role, acting as flexible linkers between folded domains. These sequences can regulate the spacing of folded domains; the conformational behaviour of a linker will intimately depend on the amino acid composition of the linker and of the protein surfaces of the two folded domains. In as of yet unpublished work [10], we developed a model to describe how a disordered protein drives a glass-transition that appears to entirely circumvent phase separation. A key part of this mechanism relies on a large, highly charged linker that ensures intermolecular interactions are strongly favoured over intramolecular interactions. In other unpublished work, we have shown how changes to the amino acid composition of the linker and of the surface residues on the two connecting domains can modulate the impact the linker has on domain-domain interaction [11]. This suggests that while disordered linkers have sometimes been implicated as passive players, they could be used to modulate polyprotein behaviour by changing the balance between *in cis* domain domain interaction and *in trans* domain interaction. Indeed, the conserved linker between two of the SH3 domains in the protein Nck plays a critical role in enhancing the formation of dynamic Nck/N-WASP/Nephrin assemblies [30]. In other unpubished work, we have shown how an enormous diversity in linker behaviour is encoded for by sequence, providing Nature with an expansive and tunable repertoire of linker behaviour [12].

[9] Greig, ..., Holehouse, *et. al* (unpublished)
[10] Boothby, ..., Holehouse, *et. al* (unpublished)
[11] Mittal, Holehouse, & Pappu (unpublished)
[12] Konig, ..., Holehouse, *et. al* (unpublished)

IDPs are also frequently found in protein and RNA chaperones, where their structural plasticity coupled with an ability to engage in extensive but weak intermolecular interactions may allow them to facilitate as local denaturants, weakening misfolded proteins and allow refolding to occur [583]. The exact mechanisms and roles of IDPs as chaperones remains unclear, and while these functions are typically considered in the context of *in trans* interactions, and intriguing hypothesis we suggest is that these regions can also act as *cis*-acting chaperones.

2.3 Experimental Methods for Studying IDPs

The absence of a well defined three-dimensional structure makes many techniques common in structural biology uninformative for providing insight into conformational behaviour of IDPs. In this section we will briefly introduce several of the experimental techniques used throughout this work.

2.3.1 Nuclear Magnetic Resonance (NMR) Spectroscopy

Nuclear Magnetic Resonance (NMR) spectroscopy has (arguably) been the most important tool for exploring and understanding IDPs *in vitro* and more recently *in vivo* [32,263,302,441, 513,570,579,646,647]. A discussion on the details of NMR (which is an entire field unto itself) is *far* beyond the scope of this thesis. Suffice to say, certain types of nuclei (which include protons and nitrogen) have an associated and detectable spin-state, and by first aligning those spins with a magnetic field, pulsing radio-waves across the sample, and observing how the

magnetization changes as a function of time, allows the NMR spectroscopist to obtain nuclei-specific information. It provides information on local chemical environments, relaxation dynamics, through bond (J-coupling) and through space (NOE) interactions, and has been used extensively to characterize many different aspects of disordered proteins, including both structural behaviour and dynamics. HSQC spectral assignments allow a convenient method for following perturbation to the local chemical environment of each residue in response to some change, such as a change to the solution environment (salt, temperature, pH), as a function of ligand or binding partner, or as a function of post-translational modifications. In chapter 6 we used NMR to obtain information on local conformational changes that emerge in response to phosphorylation. A final advantage of NMR is that it is possible to perform NMR on protein samples *in vivo* [187,512,579]. While technically challenging, in many ways this offers the holy grail of mechanistic biophysics - residue resolution insight into proteins in their (truly) native environment.

While NMR is incredibly powerful, it is also challenging. A very basic issue is cost: NMR spectrometers are incredibly expensive to purchase and maintain, so gaining access can be a major challenge. For a good signal-to-noise ratio the protein concentrations must be relatively high, which can be prohibitive for IDPs that undergo functional self association or aggregation. This concentration dependence can also be an issue for Many NMR experiments required isotope labelled protein which typically requires bacterial expression systems - this can be problematic when working with IDPs, given disordered proteins frequently express and/or purify poorly from bacterial systems. Beyond simply obtain the protein, a further challenge stems from the fact that signal is inversely proportional to the speed of solution tumbling, such that for proteins greater than ∼100 kDa the signal becomes almost *invisible*. Beyond the practical issues of obtaining NMR spectra, the data analysis and interpretation can also be highly challenge for IDPs, which are typically associated with sharp,

poorly dispersed two-dimensional spectra (i.e. highly overlapping signals due to the fact that the residues are all roughly in a similar chemical environment as a result of the proteins disordered nature).

While NMR provides high resolution insight into local conformational behaviour, for IDPs especially, it becomes more challenging to obtain global information. Pulse field gradient experiments offers a method to obtain diffusion coefficients which can be extrapolated into hydrodynamic radii using the Stokes-Einstein equation, but it remains somewhat unclear exactly how well a Stokesian assumption holds for IDPs, where fluctuations and changes in water entrainment as a function of conformation could introduce confounding factors that convolve the relationship between diffusion and average dimensions. Nevertheless, if good NMR data can be obtained for a system of interest, it will almost always be useful.

2.3.2 Small Angle X-ray Scattering (SAXS)

Small Angle X-ray Scattering (SAXS) provides insight into global conformational behaviour of proteins [183, 202, 467]. As a solution scattering approach, a concentrated protein sample is exposed to X-rays and a diffraction pattern is then collected. That diffraction pattern informs on the dimensions of the scattering particles, and although SAXS does not provide high resolution structural data, in some cases it can be used to obtain information on the shape and size of the scattering species. As a result, it has been used extensively in the characterization of IDPs [287]. The resulting scattering data can be analysed directly via Guinier analysis, fit using an ensemble of structures, or for unfolded proteins fit using a calibrated molecular form factor (MFF) to simultaneously extract information on global dimensions and the apparent solvent quality [13] [434]. In all three cases, SAXS provides

[13] J. Riback, personal communication

insight in the ensemble average global dimensions, a property that for IDPs is extremely useful for relating amino acid sequence to conformational behaviour. It also provides useful complementary information to NMR and FRET, which typically provide insight into specific local conformational behaviour. As more sophisticated methods for analysing scattering data are developed, the role of SAXS in characterising IDPs (especially in the case of larger proteins where FRET or NMR are not amenable) will likely become increasingly important.

In chapter 6, we used SAXS in conjunction with NMR and simulations to construct a holistic description of the conformational ensemble of a proline rich IDP from the yeast transcription factor Ash1. In chapter 7 we used time resolved SAXS to provide insight into the conformational behaviour of an unfolded (but foldable) protein *before* folding has begun.

As with all techniques, SAXS has various limitations. Like NMR, relatively high protein concentrations are needed, which introduces several of the same issues mentioned in the previous subsection. Unlike NMR, the issue of aggregation can be at least partially addressed using in-line size exclusion chromatography (SEC) to remove large aggregates and oligomers [483]. SAXS also has the advantage of being amenable to rapid mixing approaches, and recent advances in microfluidics have allowed such approaches to obtain scattering data in a time-resolved manner [325, 655]. Given the complexities associated with fitting scattering data, it should be no surprise that the derived radius of gyration shows an inherent dependence on assumptions made during the Guinier regime fitting [20, 59, 673]. In chapter 8, we consider the ongoing 'SAXS vs. FRET' debate, in which two techniques give apparently discrepant results for the same protein. In this chapter, we suggest that at least part of this discrepancy originates from the fact that the two methods report on fundamentally different order parameters, and while limiting homopolymer models assume these two parameters to be coupled, this need not be the case.

2.3.3 Single Molecule Förster Resonance Energy Transfer (smFRET)

Single Molecule Förester Resonance Energy Transfer (smFRET) has become an invaluable tool for the study of IDPs [71, 523]. The theoretical basis of this approach relies on the non-radiative transfer of energy between a donor dye and acceptor dye, where the transfer efficiency (the amount of energy transferred from donoro to acceptor) depends on the orientation of those two dye relative to one another and the distance between the dyes. In the limit of fast anisotropic dye rotation, the orientational component averages out, and we are left with equation 2.1

$$E(r) = \frac{R_0^6}{R_0^6 + r^6} \tag{2.1}$$

Here, $E(r)$ is the transfer efficiency, r is the distance between the two dyes and R_0 is the Förester radius, the characteristic distance when the transfer efficiency is at 50%. The transfer efficiency is measured directly as the fractional quantum yield associated with productive energy transfer from donor to acceptor, and the R_0 is a known value for different dye pairs that is measured independently. As a result, in some cases the dye-dye distance can be directly recovered, as can various other photo-physical parameters.

Equation 2.1 provides a means to convert an instantaneous transfer efficiency into a distance. However, from smFRET experiments an ensemble average transfer efficiency ($\langle E \rangle$) is measured. Therefore, a second set of mathematical tools are required to convert $\langle E \rangle$ into a distance distribution (as opposed to a single distance). This is achieved using equation 2.2

$$\langle E \rangle = \int_0^\infty E(r)P(r)dr \tag{2.2}$$

In simple terms, equation 2.2 tells us that the measured FRET efficiency is the integral associated with the FRET efficiency for each possible r distance multiplied by the probability of that distance. This is simply a continuous form of the arithmetic mean. To solve equation 2.2 requires a functional form for $P(r)$. This is typically described using one of a number of different polymer models. These models have one or more free parameters associated with them, and are fit such that these free parameter(s) lead to an instantiation of the polymer model which yield a $P(r)$ distribution that correctly reproduces $\langle E \rangle$. In this way, we can solve the inverse problem described by equation 2.2 and determine the most likely distance ($\langle r \rangle$) associated with a measured $\langle E \rangle$.

As an example, the commonly used Gaussian chain has the functional form shown in equation 2.3.

$$P(r) = 4\pi r^2 \left(\frac{3}{2\pi \langle r^2 \rangle} \right)^{\frac{3}{2}} \exp\left(\frac{-3r^2}{2\langle r^2 \rangle} \right) \tag{2.3}$$

Here, $\langle r^2 \rangle$ is the mean squared ensemble average end-to-end distance, which represents the single fitting parameter. Other models include the self-avoiding walk (SAW) model, the wormlike-chain (WLC) model and the Sanchez model. For a convenient introduction to these models in the context the recent review by Schuler, Soranno, Hofmann and Nettels is highly recommended [523].

smFRET has been critical for exploring and understanding the relationship between sequence and conformation [71, 523]. The accuracy and insight into local conformation behaviour offered by smFRET is remarkable, providing a direct measure of chain dimensions and chain dynamics [20, 59, 555, 673]. While NMR and SAXS *require* high concentrations of protein to obtain reasonable signal to noise, smFRET works best when the concentration of labelled molecules is very low. This allows the study of aggregation-prone proteins far below their saturation concentration (the concentration at which aggregation sets in) [404]. More generally, by labelling a single protein type, the heterogeneity of the surrounding environment is largely irrelevant (assuming appropriate corrections to the resulting data analysis as a function of solvent, sample depth, refractive index *etc.*). As a result, smFRET has provided incredible insight into the impact of molecular crowders on the conformational behaviour of disordered proteins [556]. Taking this a step further, a labelled sample can be delivered directly into a cell providing single-molecule insight into the conformational behaviour and chain dynamics of IDPs in their native environment [301].

Despite the power of FRET, there are a number of caveats that should be considered. Firstly, the distances being measured are not chain-chain distances, but dye-dye distances. These dyes are connected to the species of interest via a flexible linker. As a result, the influence of the linker on extrapolating chain-to-dye distances must be taken into account. The dyes typically used for single molecule experiments are fairly large planar aromatic dyes (see figure 2.6 for a sense of the relative sizes), which may introduce some bias. Computational studies suggested that these dyes have a minimal impact on the behaviour of IDPs, although given the fixed-charge nature of molecular dynamics forcefields, it remains unclear if the delocalized pi-system associated with a dye would be appropriately captured and described [665]. We caution that if, as a field, we wish to argue that IDPs show sequence-specific conformational behaviour, it seems a necessary that the nature of the dyes might lead to some dye-dye or

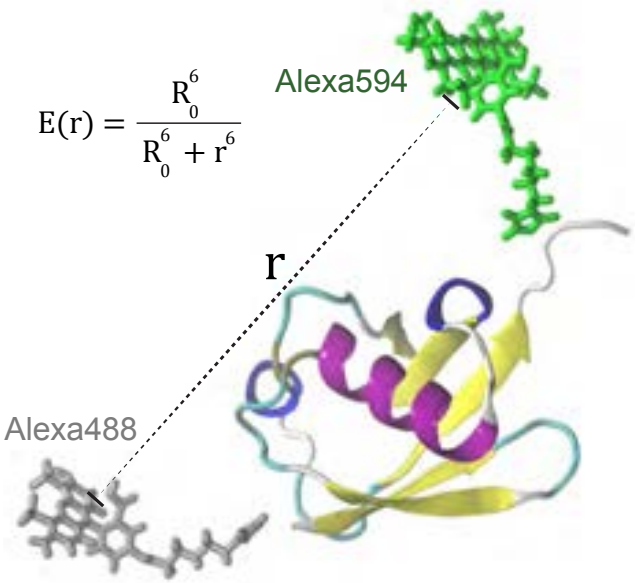

Figure 2.6: Schematic showing a dye labelled protein (ubiquitin) with the inter-dye distance (r) labelled.

chain-dye interaction, given the size, geometry, and chemistry of the dyes. In chapter 7 we use novel non-invasive FRET pairs to dissect the local conformational behaviour of the protein NTL9. Importantly, as a control, we demonstrate that for the folded state these non-invasive dyes provide a perfect description of the expected local distances based on existing crystal structures ($R = 0.99$).

A second and (arguably) more challenging issue is the conversion of transfer efficiencies to distance. As discussed, the ability to fit the measured ensemble-average transfer efficiency relies on the use of a model of P(r). These homopolymer-models introduce implicit assumptions and are necessarily describing chain dimensions in terms of the average behaviour of a single type of monomer unit. For some IDPs (especially those that behave as flexible linkers) this is a reasonable assumption, but for many others, local and long-range conformational preferences leads to anisotropic deviations from mean-field polymer models. Consequently,

while simple homopolymer models can typically be well fit to experimental data, there is not *necessarily* any guarantee that they are fitting for the right reasons, and as a result further insights obtained based on these models may be incorrect.

In chapter 8 we discuss the weaknesses associated with converting the end-to-end distance to radius of gyration, and show that depending on the extent of local structure this can grossly over- or under-estimate the global dimensions of an unfolded protein. For our work on the unfolded state of NTL9 (chapter 7), we had six independent FRET pairs distributed across the protein, providing a high-resolution global description of local conformational features. If, instead of using all six FRET pairs, a single FRET pair is used, then the extrapolated global behaviour we obtain is entirely inconsistent with the other local distances and the SAXS-derived radius of gyration. This suggests that one cannot necessarily assume that conformational behaviour derived from a single FRET pair provides a good description of global conformational behaviour. One possible solution to this is to use physics based models rather than homo-polymer theory, as we did in chapter 7 and was done in recent work by Fuertes *et al.* [189, 507]. When available, multiple FRET pairs provides a crucial self-consistency check, and as such the use of multiple FRET pairs (as has been pioneered by the Schuler group) offers a robust experimental measure of conformational behaviour across multiple chain-distances. The major drawback of this, of course, is that creating multiple independent constructs and performing the associated pairwise sets of experiments with necessarily controls is extremely labour intensive. With this in mind, a more tractable approach may be to combine one or two FRET pairs with other experimental and computational approaches (such as SAXS, NMR, or FCS), as has been done to great effect in a number of recent studies [20, 59, 189, 555, 673].

2.3.4 Fluorescence Correlation Spectroscopy (FCS)

The final technique we will briefly introduce is Fluorescence Correlation Spectroscopy (FCS) [351]. FCS was originally conceived of and developed in the 1970s by Magde, Elson and Webb, but in the last fifteen years has become a crucial component in the toolbox of experimental studying of unfolded proteins [228, 306, 539]. The basic principle of FCS is fairly simple. A species of interest (e.g. a protein) is labelled with a bright fluorescent dye and allowed to diffuse in a cuvette. A region within this cuvette is illuminated via a confocal set-up, such that as molecules diffuse into the illuminated volume they fluoresce. Many molecules diffuse in and out of the volume at each time-point, leading to fluctuations in brightness. These fluctuations are measured, and the auto-correlation function associated with these fluctuations can be fit to a simple model for a diffusing species in three dimensions. From the fit of the autocorrelation function the diffusion constant is extracted directly. The diffusion constant can provide insight in the species' global dimensions by use of the Stokes-Einstein equation. In addition to simple diffusion, more complex autocorrelation functions can be used that capture chemical reactions, species fluctuations, and a variety of other secondary and tertiary processes [163].

Like FRET, FCS allows for samples at extremely low concentration. Depending on the concentration of labelled sample, FCS may be an ensemble method (multiple species diffusing through the illuminated volume at any one time) or a single molecule methods (on average a single molecule diffusing through the confocal volume at any given time). In both cases the parameter being measured is the fluctuations in brightness. For single molecule FCS with fully labelled sample (e.g. protein of interest genetically fused to GFP) a brightness analysis can be used to determine the oligomeric state of a single species in the confocal volume [95]. In chapter 5 we used FCS to measure the length-dependent diffusion constants

associated with polyglycine, and found that, consistent with theoretical predictions and all-atom simulations, it forms compact globules. In chapter 12 we used a modified version of FCS, ultrafast scanning FCS (usFCS) to directly measure the protein concentration inside and outside of phase separated droplets.

Although FCS is extremely powerful, it typically provides less information that smFRET or NMR experiments. The conclusions drawn are also highly dependent on the models used to interpret the data, a limitation in no way limited to FCS, but in other approaches there may be additional self-consistency checks that make assessing the validity of key assumptions more straight forward. Despite this, it provides a powerful tool for examining the global conformational behaviour of disordered proteins. FCS also has the advantage of versatility - measurements can be made in 2D (e.g. on lipid bilayers) and 3D, and can be made in cells as well as *in vitro*.

2.4 Computational and Theoretical Approaches for Studying IDPs

The preceding section focused on experimental approaches for understanding IDPs. We felt it important to introduce these topics; they are tools through which the sequence determinants of conformational behaviour in IDPs can be explored. However, the majority of the work in this thesis is based on the interpretation of experimental results using a broader framework of computational and theoretical approaches. In the following sections. In the following sections we outline some general ideas in computational biophysics, followed by an overview of the

key features associated with CAMPARI-based Monte Carlo simulations and the ABSINTH implicit solvent model.

2.4.1 Introduction to Computational Biophysics

What do computational and theoretical approaches entail? In our work, the computational approaches are primarily focused on all-atom Monte Carlo simulations using the ABSINTH implicit solvent model (see work in chapters 5, 6, 7, 8, 11, and 12) [613]. In chapter 5 we use all-atom molecular dynamics simulations with explicit solvent to describe the conformational behaviour of polyglycine and two archetypal short peptides in the context of high concentrations of denaturant. In chapter 14 we introduce a novel lattice-based simulation engine for describing IDPs as simple but sequence-specific polymers. In chapters 12 and 13 we introduce a theoretical description of thermodynamics of polymer mixing to explain the complex phase behaviour of the disordered protein LAF-1. We also use a variety of sequence-based statistical analysis tools for both analysis and design of novel sequences, although we will limit our discussion of those tools to chapter 4.

Traditionally speaking, computational biophysics involves three-dimensional explicit-representation simulations of biological macromolecules. By explicit-representation, we mean that our species of interest (e.g. our protein) is represented as an entity that exist in three dimensions, although the degree of resolution is arbitrary - it could be a low as a single sphere for each protein, or high as all-atom resolution with interactions described via quantum mechanics. This is in contrast to implicit representation models, where our species are described by mean-field variables, and as such do not have any defined three dimensional topology or relative position. Such an implicit representation could be a series of coupled ordinary

differential equations to describe a signalling network, or a master-equation formalism to describe chemical kinetics. As a working definition, biophysical simulations tend to use explicit-representation models to describe complex phenomena, while systems biology tends to use implicit-representations. The use of explicit- or implicit-representation schema allow us to ask different types of questions. Importantly, both are quantitative tools for developing predictions that can be tested experimentally, and for better understanding biological systems.

An explicit-representation simulation consists of two key but independent components, a **representation scheme** and an **update scheme**.

Representation Scheme

The representation scheme is the manner in which the three-dimensional state of the system is described. By representation, we refer to the framework used to describe the relative position of the different mobile components (e.g. atoms) in the system, and also to describe how those variable components interact with one another.

In computational biophysics the representation scheme tends to be described by a **forcefield**. In the context of all-atom simulations, a forcefield is a set of rules that describe how atoms can be connected to one another (bonded terms), and how atoms interact with one another (non-bonded terms). The degree of complexity associated with a forcefield will depend on the resolution. Simple bead-spring forcefields may use a harmonic potential to connect beads together and a simple Lennard-Jones or Mie potential to dictate the non-bonded interactions [590]. Complex forcefields like AMOEBA take allow for multipole interactions

and polarization of charge distribution [470, 540]. In general, the more complex a forcefield, the more computational expensive running a simulation with that forcefield will be.

With this in mind, a molecular system can be described in terms of the actual configuration of system components (relative positions of all atoms and relative topology of molecules - i.e. what atoms are bonded to what). The energy associated with a given system-wide configuration is then determined by passing the full configuration information to the forcefield and evaluating the energy. The mapping between configuration and energy is many-to-one and deterministic; a given configuration will always give the same energy, and many different configurations may produce the same energy. In this way, we can think of a forcefield as an energy function that takes the configuration of the system as input and returns to us an energy. We will refer to this energy function-forcefield as the Hamiltonian ($\mathcal{H}$).

While early simulations were performed in a vacuum, to explore solution-state behaviour simulations are run in an aqueous environment [34, 327]. Explicit solvent simulations mean that each water molecule in the simulation is represented as an individual molecule that will interact with other water molecules and the protein in exactly the same manner as the protein interactions with itself. Having an accurate water model is critical; in an explicit solvent simulation the overwhelming majority of atoms in the system are water atoms. Consequently, if water-protein (or water-water) interactions are incorrect they can drastically bias the simulations [46, 245, 452].

Update Scheme

The update scheme is the method through which the configuration of our system evolves. In dynamics-based evolution schemes, we evaluate the change in energy upon some small

perturbation to the position of all the atoms in our system and use these changes in energy to determine an updated force (magnitude and direction) which is applied to each atom [318, 327]. In this way, Newton's equations of motion are solved. There are various different dynamic schema, including molecular dynamics, Langevin dynamics, and Brownian dynamics [318]. These differ in various aspects, but are similar inasmuch as they describe a temporal evolution of a dynamical system in which finite timesteps are taken, with the system evolving as a function of the updated atomic positions that occur over these timesteps. With the exception of the work in chapter 5, our work does not use molecular dynamics simulations, but instead uses Monte Carlo simulations, an alternative update scheme that does not use the equations of motion, but instead drives the evolution of the system through random conformational perturbations.

Monte Carlo Simulations

Monte Carlo simulations are in many ways substantially simpler to understand and implement than dynamics-based update schemes. In molecular dynamics we are allowing the system to evolve according to the forces that emerge from the interactions between all the components in the system. This means that for each timestep we update the full set of degrees of freedom in the system, and that update is (by necessity) very small. In Monte Carlo simulations we are perturbing a randomly selected single degree of freedom by a randomly determined about, evaluating the energy associated with the configuration that results from this random perturbation, and accepting or rejecting that new configuration according to some update rule [614]. For all the simulations discussed in thesis we will accept or rejected according to the Metropolis-Hastings acceptance criterion, but alternative acceptance criterion exist (e.g. Barker, Wang-Landau *etc.*) [31, 227, 498, 624] .

The Metropolis acceptance criterion is shown in equation2.4

$$p = \min\left\{1, \exp\left[-\left(\frac{1}{k_B T}\right) \times (E_B - E_A)\right]\right\} \tag{2.4}$$

Using equation 2.4 we can directly compute the probability of accepting a move as a function of kT, as shown in fig. 2.7. Moves where the energy *decreases* (becomes more favourable) are always accepted, while moves where the energy *increases* (become less favourable) are accepted with an exponentially decreasing probability. Importantly, however, even extremely unfavourable moves *can* be accepted, albeit with very low probabilities. This allows local minima to be escaped via transition states which may be substantially less energetically favourable than the associated minima.

For dynamics-based update schemes the change on each timestep is determined based on the forces experienced by each atom. For Monte Carlo simulations we must define the types of 'moves' (perturbations) that allow the system to evolve in configuration space. These moves typically involves rigid body motions (rotation and translation), as well as local moves to augment the conformation of the molecule(s) of interest. In the case of proteins, this could include dihedral angle rotation, bond stretching, ring puckering, and a range of additional processes.

For IDPs, Monte Carlo simulations offer a distinct advantage over molecular dynamics simulations. Molecular dynamics algorithms (by design) construct a smooth trajectory whereby the system evolves along a energy-gradient dictated by a combination of the forces experienced by each atom and the thermal fluctuations provided by the surrounding environment. Consequently, upon reaching a local meta-stable minimum, if the fluctuations experienced by the macromolecule are small enough relative to depth of that minima, a molecules may

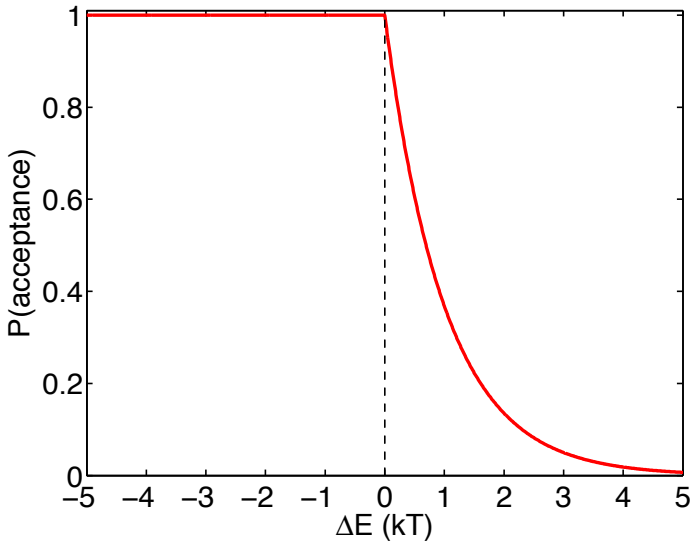

Figure 2.7: Plot of acceptance probability vs. change in energy associated with a move. Moves that lead to a decrease in energy will always be accepted, while moves that increase energy may be accepted, depending on how significant the increase in energy is. There is always a finite chance of accepting a move, such that for moves that lead to an increase in energy of 5 kT there is a 0.67% chance of acceptance; low but not zero.

become 'trapped' in a local meta-stable state for a significant length of time. This is not an artefact, but reflects the true depth of the minima in the context of the free energy landscape. Unfortunately, our simulations run for a finite amount of time, and remaining trapped in a local minima may lead to a gross under-sampling of the available conformational space, effectively breaking the ergodic assumption.

To illustrate this point, we can construct a hypothetical energy surface associated with an IDP, as depicted in fig. 2.8. Here, the vertical axis corresponds to energy, and the horizontal axes are some arbitrary description of conformation. The black spheres (starting at the local minimum labelled A) represent the path associated with some hypothetical molecular

dynamics trajectory. The trajectory explores along the gradients of energy, spending the majority of its time sub-sampling states around A. The white spheres (also starting at the local minimum labelled A) represent the path associated with some hypothetical Monte Carlo trajectory. This trajectory is undergoes a series of jumps to explore many different minima, a behaviour possible through large-scale reconfigurations that allow the system to directly reconfigure over the energy barriers that prevent the equivalent sampling by molecular dynamics simulations.

For folded proteins this is (to some extent) less of an issue; assuming the simulation begins with a protein in its folded state, we are typically only interested in exploring *within* that folded basin. This folded basin can encompass the range of local conformational dynamics associated with folded proteins. For IDPs, however, the free energy surface that describes a disorder protein is (almost by definition) extremely rugged. For disordered proteins where the many local minima that give rise to this rugged surface are shallow (relative to kT) this is not a major problem, as escape from any given minima is likely. However, for IDPs where these local minima are deeper, the efficiency of an MD simulation will decrease exponentially as a function of well depth and number of wells. Consequently, despite running long molecular dynamics simulations, an IDP may only samples a handful of distinct states, with the majority of the time spent engaging in small local re-arrangements as the protein explores a single local minimum. Monte Carlo simulations, on the other hand, are able to jump between local minima through single moves that bypass high energy barriers or allow transient passage over such a barrier via the unlikely (but possible) acceptance of a high energy move. By designing move-sets that maintain detailed balance but allow for large-scale re-arrangements in a single move, we can explore conformational space in a much more efficient manner [614].

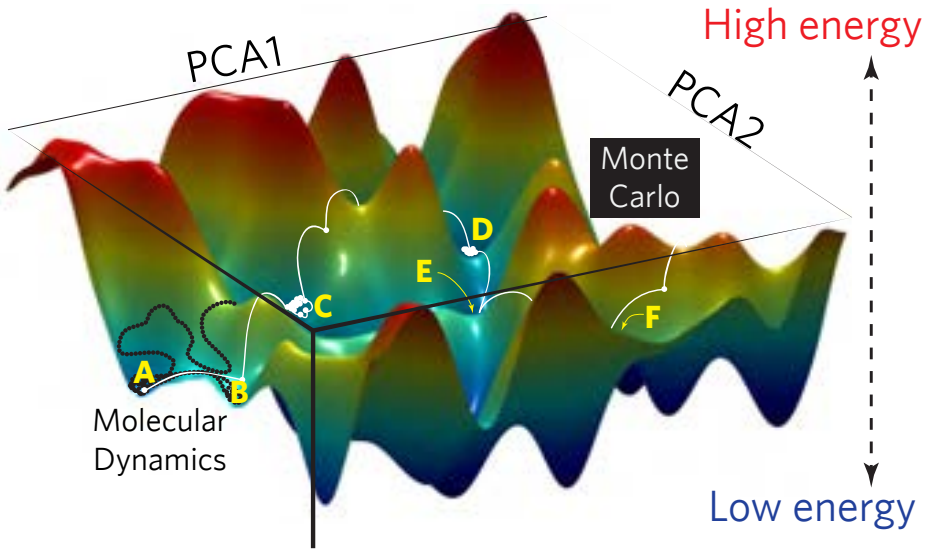

Figure 2.8: Schematic examining a hypothetical free energy surface of an IDP and two putative trajectory pathways over that surface. In both cases these trajectories begin at state A. The **black path** is drawn to represent the trajectory associated with a molecular dynamics simulation. Each configuration is equidistant from the previous (in phase space) due to well defined finite steps, and the majority of the simulation is spent sampling the minima at A and B, eventually crossing the larger barrier towards state C. The **white path** is drawn to represent the trajectory of a Monte Carlo simulation. Large 'jumps' between different configurations are possible due to moves that cause large-scale conformational changes, allowing the system to entirely bypass the majority of the high barriers and explore various states (A-through-F). Naturally this is a biased schematic drawn to deliberately suggest that Monte Carlo simulations as highly efficient, but the point remains that these jumps in phase space allow Monte Carlo simulations navigate an inverted egg-box style free energy landscape (one with many minima of approximately the same stability).

More generally, a fundamental challenge in the simulations of IDPs is the inability to assess the true size of conformational space accessible to the protein of interest. An MD simulation

run for 5 µs may appear well sampled, but assessing the true extent to which the simulation is exploring conformational space is almost impossible. Put another way, what fraction of the possible states accessible to an IDP are being found? There is no obvious way to know this without complete enumeration with the appropriate Hamiltonian, at which point your have already fully sampled the system. Qualitative litmus tests to asses sampling including running multiple independent simulations or assessing how derived results change if various subsets of data are used, but these are far from rigorous. In the seminal review by Zuckerman and Grossfield, the sentence, "*Visual confirmation of good sampling is still an important check on any quantitative measure*" is as accurate as it is alarming [207]. These challenges apply to both Monte Carlo and molecular dynamics simulations, but the inherent nature of Monte Carlo simulations makes them well suited for exploring rugged/non-convex energy surfaces.

In chapter 9 we introduce an algorithm designed (in part) to assess the degree of local conformational sampling. Such an approach can be useful for providing a quantitative framework for thinking about how well converged simulations are, but will typically be most useful in the cases of pathologically poorly sampled simulations. In chapter 14 we will briefly touch on a new class of Monte Carlo moves for sampling rugged energy landscapes (Temperature Sweep Metropolis Monte Carlo). To be clear, we do not mean to imply that Monte Carlo simulations are not subject to exactly the same sampling challenges as molecular dynamics simulations are. These are especially true in the case of deep local minima where single moves that allow escape are not available. In light of this, enhanced sampling approaches including temperature replica exchange, Hamiltonian replica exchange, and Hamiltonian Switch Metropolis Monte Carlo play key roles in hard-to-sample systems for Monte Carlo simulations [394, 565, 656].

A drawback of Monte Carlo simulations (compared to molecular dynamics simulations) is that unless explicitly encoded in the implementation of the move set, Monte Carlo simulations do not generate a trajectory that provides useful kinetic information [175]. Each move involves a randomly selected degree of freedom and a random extent of perturbation. As a result some steps may lead to tiny changes in the system's conformation, while others may lead to large conformational re-arrangements, meaning the interpretation of reconfiguration vs. number of steps depends on the moves proposed.

Given that Monte Carlo simulations evolve via an accept/reject mechanism, a major advantage is the fact that there is no need to calculate the forces associated with each atom. Moreover, because a single degree of freedom is being perturbed at a time, intelligent algorithms can be designed that only evaluate the energy associated with the changing components in the system. As a result, Monte Carlo simulations can be incredibly fast, although they are often less amenable to parallelization (unlike molecular dynamics, where domain decomposition and GPUs have been transformational in improving wall-clock time) [1, 275, 294, 514].

For explicit solvent all-atom simulations, Monte Carlo simulations suffer from one major drawback: because the moves involve the perturbation of single degrees of freedom in an independent manner, in dense systems the majority of moves are rejected due to steric clashes. Given the density of liquid water, this is an issue for all-atom simulations with explicit water under aqueous conditions. Dense systems need not necessarily be prohibitive for Monte Carlo simulations (in chapter 14 we report results from simulations where the volume fraction is up to 90%), but nevertheless typically lead to a large increase in the number of steps required to obtain converged results due to the plummeting of the acceptance ratio (the fraction of proposed moves which are accepted).

To alleviate these density-induced inefficiencies, the Monte Carlo simulations performed in this work take advantage of an implicit solvent models [613]. In an implicit solvent models, instead of representing each water molecule as a separate species in the simulation, the water is treated as a mean-field interaction, where the strength of the associated solvation energy is related to the solvent accessible volume associated with the protein surface. There are various possible approaches to implement an implicit solvent model. For our all-atom simulations we used the ABSINTH implicit solvent model. In chapter 14 we introduce a new coarse-grained forcefield, the General Chemical Forcefield (GCF) forcefield, which while still in development provides promising early results at a fraction of the cost.

2.4.2 CAMPARI

The CAMPARI simulation engine is used extensively throughout this thesis. CAMPARI is a feature-rich, powerful simulation engine for performing molecular dynamics (in both Cartesian space and torsional space) as well as for performing Monte Carlo simulations [613–615]. In this subsection we will focus on its capacity as a Monte Carlo simulation engine, and for the body of work described herein all simulations performed with CAMPARI are Monte Carlo simulations. The specific details associated with CAMPARI are well documented online (http://campari.sourceforge.net/) and will not be repeated here. In short, the degrees of freedom for conformational sampling are illustrated in table 2.4.2

Perturbations to these degrees of freedom are of two types: fixed perturbations, that alter a degree of freedom by a specific step-size, or randomized perturbations that first pick a random extent of perturbation (a numerical value selected from a uniform distribution of

Location	Degree of freedom
Molecule	Rigid body coordinate (position and orientation)
Backbone	ω angle ($CA_{i\text{-}1}$, $C_{i\text{-}1}$, N_i, CA_i)
	ϕ angle ($C_{i\text{-}1}$, N_i, CA_i, C_i)
	ψ angle (N_i, CA_i, C_i, N_{i+1})
	Proline (has seven non-redundant degrees of freedom to facilitate puckering)
Sidechain	Depending on residue has ≥ 0 χ_1, χ_2, χ_3, χ_4 angles

Table 2.1: Degrees of freedom in the CAMPARI Monte Carlo simulations. Note i here reflects the index of a specific amino acid.

values between 0 and some pre-defined upper bound) and then alter the degree of freedom by this randomly selected value.

In CAMPARI Monte Carlo simulations bond lengths and angles are held fixed, a treatment which is used frequently for Monte Carlo simulations of biomolecules and does not introduce artefacts, although this statement does not holds true for molecular dynamics simulations [445, 446]. By holding bond lengths and angles fixed the effective phase space accessible to the simulation is reduced, but this reduction is of predominantly non-relevant conformations, in effective providing a substantial improvement in sampling. We can consider this to be equivalent to rejecting all moves that perturb bond angles and bond lengths that deviate from the ideal value. For molecular dynamics, such stiff bond angles and lengths would cause local barriers to dynamics, but given Monte Carlo simulations evolve through random perturbations to the degrees of freedom such barriers are not an issue.

Simulations are performed in a spherical environment with a soft-wall boundary potential that has a radius of typically 2-3 times the contour length of the polypeptide. Finite size

effects are tested for by running simulations with $\mathcal{H}_{\mathrm{EV}}$ (discussed in subsection 2.4.3) at several different sizes and identifying the droplet radius at which end-to-end chain compaction is experienced.

Long range electrostatics are computed by explicitly dealing with all monopole-dipole and monopole-monopole interactions at atomistic resolution (keyword `FMCSC_LREL_MC 1`). CAMPARI allows for several distinct approaches for computing long-range electrostatics, but we mention this choice as it deviates from the default behaviour (`FMCSC_LREL_MC 3`). Beyond this choice, all options used for simulations are as associated with a default keyfile. The default values defined within CAMPARI can change; as such, we recommend setting all critical keywords explicitly in the keyfile to ensure that no hidden surprises emerge.

2.4.3 ABSINTH

The ABSINTH implicit solvent model provides a way to evaluate the instantaneous energy associated with a three-dimensional configuration of some biomacromolecule. In this capacity, it allows a software package (in our cases CAMPARI, but in principle any simulation engine) to perform all-atom simulations (either Monte Carlo or molecular dynamics) by evaluating the energy associated with a given state. The practical mechanism by which the energy is evaluated is an implementation detail of the software in question. Instead, we can think of ABSINTH as an analytical description of energy as a function of atomic position and topology. In this way, the distinction between ABSINTH and CAMPARI should be clear.

ABSINTH is different from other implicit solvent models. In ABSINTH, biomacromolecules are decomposed into a set of solvation groups based on the molecule's chemistry. Experimentally determined free energy of solvation information are available for each solvation group. In this manner, the impact of solute-solvent interaction is captured directly by combining experimental data with a functional relationship between the solvent accessible volume and the solvation state of each atom. This mapping of solvent accessible volume to solvation state uses a stretched sigmoidal function, providing an analytical mapping that provides an approximation for the partial solvation of atoms.

The total energy associated with a given conformation of a macromolecule is defined as

$$E_{total} = W_{solv} + W_{el} + U_{LJ} + U_{corr} \tag{2.5}$$

As can be seen, the ABSINTH model consists of four distinct terms: a solvation term W_{solv}, an electrostatics term W_{el}, a Lennard-Jones term W_{LJ} and a torsional correction term U_{corr}.

Mean Field Solvation Term (W_{LJ})

The mean field solvation term describes how the solute interacts with the solution environment. As described, the solute (e.g. protein, nucleic acid, ion, *etc.*) is decomposed into non-overlapping **solvation groups**, where each atom associated with the solute belongs to exactly one solvation group. These solvation groups represent distinct chemical moieties for which experimental data regarding the free energy of solvation have been measured [72].

W_{solv} is written as

$$W_{solv} = \sum_{i=1}^{N_{SG}} \sum_{k=1}^{n_i} \lambda_{i,k} \nu_{i,k}^{\text{solv}} \Delta G_i^{\text{solv}} \tag{2.6}$$

Here, N_{SG} is the total number of solvation groups in the system, n_i is the number of atoms in solvation group i, $\lambda_{i,k}$ is a weighting factor for the kth atom of solvation group i that lies between 0 and 1, $\nu_{i,k}$ is the solvation state of the kth atom in group i (again, which lies between 0 and 1), and ΔG_i^{solv} is the experimentally determined free energy of solvation associated with that group. $\nu_{i,k}^{\text{solv}}$ is related to the solvent accessible volume fraction by a stretched sigmoidal function (i.e. when the atom is fully solvent accessible $\nu_{i,k}^{\text{solv}} = 1.0$ and when it is fully solvent inaccessible $\nu_{i,k}^{\text{solv}} = 0.0$). In the standard ABSINTH implementation the $\lambda_{i,k}$ values are distributed based on an analysis of the changes to the free energy of solvation associated with a series of relevant homologous compounds. As an example, in the homologous series formamide (CH_3NO), acetamide (C_2H_5NO), propionamide (C_3H_7NO), and butaneamide (C_4H_9NO) the addition of sequential methylene groups has a negligible impact on the free energy of solvation; consequently the λ weights associated with those groups are also negligible.

Electrostatics Term (W_{el})

The term W_{el} describes the polar interactions within the system. Here, polar interactions refer to interactions between fully-charged moieties (e.g. ions or carboxylate groups), and between partial charges. All atoms contain some degree of partial charge, while only a small subset are part of a chemical group that holds a fixed integer charge. For atoms that originate from net-neutral chemical groups a distance cut-off is applied, but for fully charged groups no distance cut-off is used (although the natural form the Coulomb potential

means the effective interaction beyond some distance is minimal). As a result, while the computational cost-per-step of CAMPARI+ABSINTH simulations scales reasonably well in response to number of amino acids, it scales more poorly as a function of ion concentration.

In much the same way as solutes are broken down into solvation groups for the W_{solv} term, solutes are also broken down into **charge groups**. Charge groups represent collections of atoms that each posses a partial charge but for uncharged groups the sum of the partial charges associated with those atoms is net neutral. For groups with a fixed charge (ions, carboxylate groups, *etc.*) the summed partial charges add up to an integer charge. The membership and topology of charge groups is not the same as solvation groups, but are instead based on the partial charge assignments associated with standard molecular mechanics forcefields. For ABSINTH we typically use either OPLS-AA or CHARMM based partial charges, meaning ABSINTH simulations are in fact done using ABSINTH-OPLS or ABSINTH-CHARMM. As in standard molecular mechanics forcefields, charge-charge interactions only occur between atoms in distinct charge groups, but not within the same charge group.

For all the work in this thesis we used ABINSTH-OPLS, but when tested identical or approximately identical results were obtained with ABSINTH-CHARMM for most systems (see below). ABSINTH-CHARMM has the advantage of including a broader repertoire of parameters (including phosphorylated sidechains and charge-neutralized side chains). However, ABSINTH-OPLS includes a proline-specific parameter set that accurately reproduces *cis-trans* statistics for polyproline [471]. For our work in chapter 6 on the proline-rich IDP from Ash1 we found that ABSINTH-CHARMM was unable to reproduce SAXS data, while ABSINTH-OPLS reproduced it well. We provide these anecdotal results to warn future

users that proline-rich sequences *may* be better reproduced using the updated ABSINTH-OPLS (with the parameters of (Radhakrishnan *et al.*), as is standard in the parameter set associated with `abs3.2_opls.prm`. To help relieve this issue the integration of the proline parameters into the ABSINTH-CHARMM forcefield is natural next step.

The W_{el} is written as

$$W_{el} = \sum_{i=1}^{N_{CG}} \sum_{k=1}^{n_i} \sum_{j=i+1}^{N_{CG}} \sum_{l=1}^{n_j} f_{[(i,k):(j,l)]} \left(\frac{q_{(i,k)} q_{(j,l)}}{4\pi\epsilon_0 r_{[(i,k):(j,l)]}} \right) s_{[(i,k):(j,l)]} \tag{2.7}$$

This provides a summation over all unique pairs of atoms in distinct charge groups. Here, N_{CG} is the number of charge groups in the system, and we use the notation $[(i,k):(j,l)]$ to refer to atom k from charge group i and atom l from charge group j.

$f_{[(i,k):(j,l)]}$ is effectively and on/off switch, and is set to 0 if the two atoms in question are connected via a direct covalent bond, are part of a bond angle, or are part of the same charge group, otherwise the value is set to 1. $q_{(i,k)}$ is the partial charge associated with atom k in charge group k. $r_{[(i,k):(j,l)]}$ is the distance between atom k from charge group i and atom l from charge group j. ϵ_0 is the permittivity of free space. Finally $s_{[(i,k):(j,l)]}$ represents a correction factor that accounts for inhomogeneities in the mean-field dielectric caused by many body-effects that modulate the solvent accessible volume associated with the two atoms of interest, and is defined as

$$s_{[(i,k):(j,l)]} = \left[1 - a\nu^{\mathrm{el}}_{(i,k)}\right]\left[1 - a\nu^{\mathrm{el}}_{(j,l)}\right] \tag{2.8}$$

and

$$a = 1 - \frac{1}{\sqrt{\epsilon}} \tag{2.9}$$

Like the solvation state $\nu^{\text{solv}}_{(i,k)}$, the parameter $\nu^{\text{el}}_{(j,l)}$ is the electrostatic-solvation state. This is also defined by a stretched exponential, but is not the same as the standard solvation state. In other words, there are two distinct measures of solvation state associated with each atom ($\nu^{\text{solv}}_{(i,k)}$ and $\nu^{\text{el}}_{(j,l)}$) - one is used for W_{solv} and the other for W_{el}.

Lennard-Jones (Short Range) Interaction Term (U_{LJ}) and U_{corr}

Like many other molecular mechanics forcefields, ABSINTH uses a 12-6 Lennard-Jones potential to describe closest approach interactions. The Lennard-Jones terms used for ABSINTH are different from those in other forcefield packages, and have been meticulously calibrated to reproduce the heats of fusion and densities of small molecule crystals [593,614]. As a result, the ABSINTH Lennard-Jones parameters typically give rise to smaller hard-sphere radii than are obtained in other forcefields; these parameters allow for accurate reproduction of various physical phenomena, and have shown good agreement with various models. Moreover, these LJ parameters typically allow ABSINTH simulations to produce Ramachandran statistics that match experimental data much more accurately than standard molecular mechanics forcefields without various corrections (e.g. CMAP).

As is standard, the functional form the 12-16 Lennard-Jones potential is written as

$$U_{LJ} = 4 \sum_{i}^{N} \sum_{i+1}^{N} f_{i,j} \epsilon_{i,j} \left[\left(\frac{\sigma_{i,j}}{r_{i,j}} \right)^{12} - \left(\frac{\sigma_{i,j}}{r_{i,j}} \right)^{6} \right] \tag{2.10}$$

Where N is the total number of atoms in the system (such that i and j are distinct atoms), $f_{i,j}$ is set to 1 if atoms i and j are separated by one or more rotatable bonds, else it is set to 0, $r_{i,j}$ is the distance between atoms i and j, and the parameters $\epsilon_{i,j}$ and $\sigma_{i,j}$ are the characteristic interaction parameters for the interaction between atoms i and j.

Finally, U_{corr} is a catch-all term that includes several geometrical potentials to maintain planarity of certain bonded systems, such as amide bonds and the hydroxyl group associated with tyrosine. These corrections capture local stereospecific electronic effects that would not be captured by local steric effects.

Final Comments

ABSINTH+CAMPARI is a remarkably powerful tool. Despite the fact that solvent is represented as a mean-field interaction, it has repeatedly been able to accurately capture the conformational behaviour of a wide range of disordered proteins, often for systems where conventional simulations were explicitly tested and failed [200, 381]. Conventional all-atom explicit solvent approaches have typically suffered from three key challenges when simulating disordered proteins;

1. An enrichment for secondary structure was frequently observed in older forcefields. Indeed, not only an issue for IDPs, this was a challenge for protein folding studies [186]]. This has largely been corrected with newer versions of forcefields.

2. A tendency to over-compact is also generally observed for many explicit-solvent all-atom forcefields. This may in part be due to over-zealous protein-protein interactions, which have to some degree been parameterized (or at least evaluated against) folded protein structures. However, a growing body of evidence suggests that protein-water

interactions may also be to blame, with several recent papers tweaking the solvent-solute interaction strengths to reduce collapse [44, 46, 245, 452, 490]. These forcefields show promising results, although the extent to which both folded behaviour and the full range of disordered proteins are correctly capture with the requisite sequence specificity remains unclear.

3. As described above, the more fundamental challenge for IDPs is the problem of sampling. In this regard, approaches such as PIGGS, FAST and other equilibrium-based enhanced sampling tools may be well suited to rapidly explore conformational space [22,676]. Despite this, the compounding challenges of the computational cost of explicit solvent and the inherent slowlness for the exploration of configuration space may make explicit solvent simulations prohibitive for obtaining converged well sampled ensembles for some time to come. Two anonymous anecdotes of interest (1) For an IDP of around 30 residues it took just under a month for replica-exchange explicit solvent simulations to provide a converged and accurate ensemble, while it took a single temperature run with ABSINTH and CAMPARI around six hours to achieve almost indistinguishable accuracy. (2) I spoke once with an individual who had a colleague who had been running simulations of α-synuclein with explicit solvent for $\sim$1.5 years. These were yet to equilibrate. These are perhaps extreme examples, but help illustrate a more general challenge in obtaining accurate ensembles for IDPs.

2.4.4 IDPs and Analytical Theory

While simulations have been used extensively in the world of disordered proteins, an accurate but fully analytical description lags behind. Recent work by Sawle & Ghosh and by Lin & Chan provide elegant descriptions of how analytical theory can be used to predict

conformational and emergent properties, although these theories are limited to idealized systems [335–337, 517]. In the case of the work by Lin & Chan, progress towards a general framework for understanding the determinants of phase separation in heteropolymeric systems has been made, although fully mean-field theories such as the random-phase approximation necessarily lead to a loss of specific types of interactions. Our work in chapters 12 and 13 suggests this could be an issue. While an effective theory for the prediction of sequence-specific conformational properties would be incredibly powerful, it remains unclear if such a theory would have any advantages over numerical simulations, especially if those simulations could be done with simple models and yield predictive and at least qualitatively correct results (see 14).

Despite the fact that IDPs are (generally) heteropolymers, there is a wealth of polymer physics based on homopolymers we can leverage to understand IDPs [67, 234, 440]. Many of these topics are considered in the introductory chapter 3, and again in chapters 5, 7, 12 and 13.

One concept which we will return to is in chapter 5 is the effective scaling exponent ν. We can consider a polymer in terms of the balance between chain-chain and chain-solvent interactions. In this framework, the scaling exponent ν provides a route to predict the dimensions of a polymer as a function of the number of monomers in the chain. This relationship is captured by the expression;

$$\langle \text{dimensions} \rangle = R_0 N^{\nu} \tag{2.11}$$

Where R_0 is a prefactor scalar that captures a combination of the the bulk of the monomer and the chain persistence length.

If chain-solvent interactions are preferred over chain-chain interactions the polymer is said to be in a 'good' solvent, and $\nu > 0.5$. In the limit of an infinitely long polymers, a chain in a good solvent will necessarily reproduce the global dimensions expected for a self-avoiding random chain and $\nu = 0.588$. If chain-chain and chain-solvent interactions are perfectly counter-balanced the polymer is said to be in a Θ solvent and $\nu = 0.5$. A chain in a Θ solvent is also referred to as Flory Random Coil, sometimes just a random coil[14], a Gaussian chain, and a random flight chain. When chain-chain interactions are preferred over chain-solvent interactions the polymer is said to be in a 'poor' solvent and $\nu < 0.5$. For an an infinitely long polymer, a chain in a poor solvent forms a dense globule and $\nu = 0.33$, a state sometimes referred to as a an equilibrium globule or compact globule. As an aside, while polymers in a good or Θ solvent have equivalent global and local scaling (i.e. show fractal dimensional behaviour), a compact globule does not. However, the (appropriately named) fractal globule does show both $\nu \approx 1/3$ scaling and shows fractal behaviour, but is a fundamentally non-equilibrium state [662].

The stationary points for ν (0.33, 0.5, and 0.5888) reflect limiting behaviours for infinitely long chains. For real chains, intermediate values of ν can be achieved, although we should be cautious of how these values are interpreted. For a finite-length polymer, there is (by definition) a length dependence associated with ν, such that we must consider these intermediate values of ν to be ν^{app}. This should not be taken to mean that intermediate values of ν^{app} are not useful, as rather than describing scaling behaviour they can (along with the prefactor R_0) provide direct quantitative insight into the chain-solvent interactions and deviations from expected polymer models. A second issue is that scaling behaviour is inherently a property that makes sense for a homopolymer, but for a heteropolymer the impact local

[14]Beware - *random coil* is a loaded term and means different things to structural biologists, polymer physicists, and NMR spectroscopists. You have been warned.

interactions between specific chemical moieties could lead to conformational behaviour that may be inconsistent with simple scaling theories on some length-scales.

Using all-atom resolution models we can use CAMPARI and ABSINTH to generate conformational ensembles that match these three scaling limits. This allows a well posed question to be asked: *If* an IDP of interest behaved as a homo-polymer in the *true* poor, Θ or good solvent regime, what should our expectations be in terms of the associated conformation ensemble? We suggest this is a more interpretable way to think about solvent quality than directly fitting for a scaling exponent, given the length dependencies discussed above.

To generate these ensembles we take advantage of a modified Hamiltonian.

1. For polymers in a poor solvent we use a modified Hamiltonian - $\mathcal{H}_{\mathrm{LJ}}$ - where the only terms are the attractive and repulsive terms of the Lennard-Jones potential. The repulsive part prevents steric overlap, while the attractive part makes all atoms effectively (close to) uniformly sticky for one another. As a result, non-specific globules form and provide a good reference for poor solvent behaviour. Of note, the quality of sampling in the poor-solvent regime is a major issue, such that to generate useful ensembles the best approach is to run many (¿500) extremely short simulations and construct a 'meta' ensemble from these.

2. For polymers in a Θ solvent we use a modified Hamiltonian - $\mathcal{H}_{\mathrm{FRC}}$ - where *all* attractive and repulsive terms are turned off, creating a phantom chain. To ensure we create an ensemble of locally reasonable states, rather than use a standard CAMPARI simulation that samples the degrees of freedom listed in table 2.4.2 we instead use the Flory rotational isomer approximation, and randomly set local amino acid dihedral angles

based on a library of allowed values [177]. This approach requires a modified version of CAMPARI, and should not be attempted with the standard implementation.

3. For polymers in a good solvent, we again use a modified Hamiltonian - $\mathcal{H}_{\text{EV}}$ - where only the repulsive terms of the Lennard-Jones potential are used. This means that only interactions between atoms prevent steric overlap.

These limiting cases are not representative of real IDP behaviour, but they provide immensely convenient sequence-specific reference states for normalizing against the effects of chain topology. We have used these reference states in several of the analyses deployed throughout this thesis. The reference states help set our expectations, and provide relativity for results relating to long-range, local, and global chain dimensions.

As we discuss in chapter 7, the analysis associated with scaling exponents was developed in the limit of homopolymers consisting of millions of monomers. Should such an analytical framework be predictive (or even relevant) for finite-length heteropolymers? In the immortal words of Mr. Linch, "*Kind of, but not really*". ν is a convenient order parameter, and provides some measure of a chain's ensemble average global behaviour. That said, this utility should not be assumed to extend to local interactions, or indeed necessarily provide predictive insight into inter-molecular vs. intra-molecular behaviour. We demonstrate in 7, ensembles that give rise to an identical values for ν can have a wide range of global and local conformational properties, and show in chapter 13 that IDPs can entirely decouple their inter-molecular and intra-molecular interactions.

We finish this section with a general sentiment that will echo throughout the first half of this thesis. We have gained incredible mileage using physics that was developed to describe homopolymers and applying it to proteins. For some proteins (e.g. folded proteins under

strongly denaturing conditions) these models provide a quantitative description of the global and local behaviour [297]. In the case of strongly denatured proteins, despite the chemical complexity associated with the sidechains, in the limit of a high concentration of denaturant we have effectively converted the heteropolymer into a homopolymer from the perspective of chain-chain and chain-solvent interactions; hence the robust $\nu = 0.59$ scaling.

However under normal aqueous conditions the chemical complexity presented by protein sidechains pulls local conformational behaviour away from the manifold described by simple homopolymer models in a strongly anisotropic manner. As a result, we should expect these models to be progressively less useful as further high-resolition studies of local conformational emerge. It is at least plausible that this fact in isolation qualitatively explains the apparent discrepancy between SAXS and FRET at low denaturant concentration. In chapter 8 we will show that it also does so quantitatively. For folded proteins, we may find it easier to rationalize this result - after all, it should be expected that as well defined secondary and tertiary structure begins to emerge the ability of mean-field homopolymer models to describe those local conformational preferences should deteriorate rapidly. However, a *key* message we wish to convey in this work is that those same, well-defined anisotropic interactions that we observe during the early stages of protein folding that are also present in IDPs. IDPs are not Gaussian chains, but show strong, sequence dependent conformational behaviour across many length-scales. This does not mean that their global behaviour *cannot* be described by a polymer of the same number of monomers in some theoretical scaling regime (and in fact it is almost impossible for this not to be true), but this should *not* be treated as proof that they are devoid of local conformational preferences. The bottom line is this; homopolymer models have provide huge insight, but this does not mean IDPs are homopolymers.

2.5 Final Remarks

We conclude this chapter with a brief summary of the key ideas. We have introduced intrinsically disordered proteins and provided some rational for their prolonged absence from textbooks. We then considered the relationship between amino acid sequence and conformational behaviour, a topic that makes up the majority of the material in part I. We then discussed the function and evolution and IDPs - although not topics explored in this work we feel it important to introduce them to provides some broader scope for the interplay between sequence and phenotype. We then introduced experimental and computational methods for studying IDPs, with an extended discussion on the ABSINTH implicit solvent model, which will be used extensively throughout this work. Finally, we have touched on some of the concepts (and challenges) associated with relating sequence to conformational behaviour via analytical theory.

Chapter 3

Phase separation in biology

The final introductory chapter in this work describes the topic of phase separation in biology. We will briefly discuss some of the associated physics, although we defer to the extensive material in in chapter 13 with respect to the thermodynamics of polymer mixing. We will then discuss the challenges of naming this phenomena, and briefly consider gels vs. liquids. Next, we will discuss the types of molecules that facilitate biological phase separation, and then consider the ultimate question: *why* might Nature be using phase separation at all. Finally, we will consider how different types of amino acids may drive phase separation in different ways.

3.1 An Introduction to the Physics of Phase Separation

Everyone loves salad dressing. Some of us love it for bringing a flash a taste to a distinctively herbivorous dish. Others love it because it provides an ideal pedagogical framework with which to introduce the concept of liquid-liquid phase separation. In the interest of simplicity

(and to the disdain of the gourmand) let us imagine a simple salad dressing consisting only of oil and vinegar. Both oil and vinegar are liquids; they flow, drip, wet a surface, will deform to occupy the volume of their encompassing container to a limit dictated by surface tension, and on a molecular level experience rapid internal reorganization. This does not mean there can't be preferential interactions within a liquid - as a prime example, water shows a distinct radial distribution profile due to preferential intermolecular hydrogen bonding (see fig. 3.1) [413]. However, in spite of any well defined local interactions, given infinite time, a liquid will deform and flow. Essentially, the strength of those preferential interactions are unable to overcome the entropic driving force that gives rise to internal re-arrangement, but strong enough to create a cohesive network, leading to a surface tension.

Returning to our delicious analogy, both oil and vinegar are liquids, but if we mix them together and then allow them to stand for a little time something curious occurs. Despite the fact both are liquids, we find that they will **demix** into distinct droplets of vinegar in a sea of oil (or *vice versa*, depending on the ratio of vinegar:oil). Figure 3.2 illustrates this phenomenon. To most people this is simply a mildly inconvenience, but over the last ten years there has been an emerging consensus that the physics that underlies this behaviour represents a critical mechanism through which Nature facilitates cellular organization [27].

What is going on in these droplets? Inside the vinegar droplets, the concentration of vinegar is high and the concentration of oil is low (but still finite). Similarly, outside of the vinegar droplets the concentration of vinegar is low (but still present) but the concentration of oil is high. These two phases - inside the droplet and outside the droplet - are described from a vinegar-centric perspective as the dense phase and the dilute phase, respectively[15]. Both phases are dynamic, undergoing rapid internal re-arrangement consistent with a liquid. In

[15]From the oil's perspective, the vinegar rich droplets are the dilute phase and the enveloping solution is the dense phase

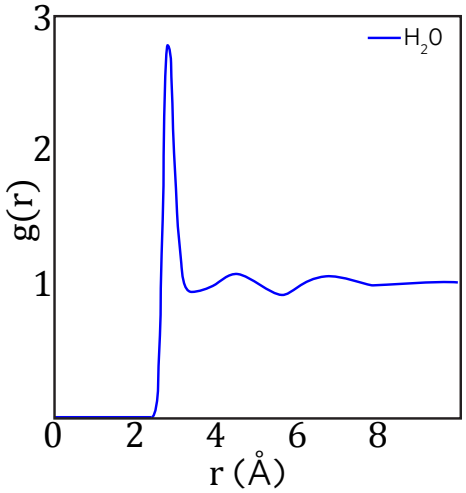

Figure 3.1: Schematic of a water pair correlation function. The pair correlation function can be thought of as the radial probability of encountering another molecule of interest (in this case water), normalized for the radially growing volume element. The area under the first peak provides insight into the number of water molecules found in the immediate vicinity, which for liquid water is ≈ 4.4 and decreases to exactly 4 in ice. The fact that the peaks decay to zero beyond some distance does not (necessarily) mean water experiences no long-range interactions, but only that over these longer ranges any interactions are isotropic.

addition to this internal rearrangement, while there is no net flux between the two phases, there is a constant steady-state exchange of oil and vinegar shuttling between the two phases in a dynamic equilibrium. On average, the same number of vinegar molecules are in the dense phase at any given time, but the identity of those molecules is constantly changing.

A convenient experimental method to quantify this dynamic behaviour is to use Fluorescence Recovery After Photobleaching (FRAP) [479]. In this approach, our species of interest (e.g. the vinegar) is tagged with a fluorescence label such that all the molecules fluoresce; consequently, the droplets appear as bright foci under an appropriate light source. If the

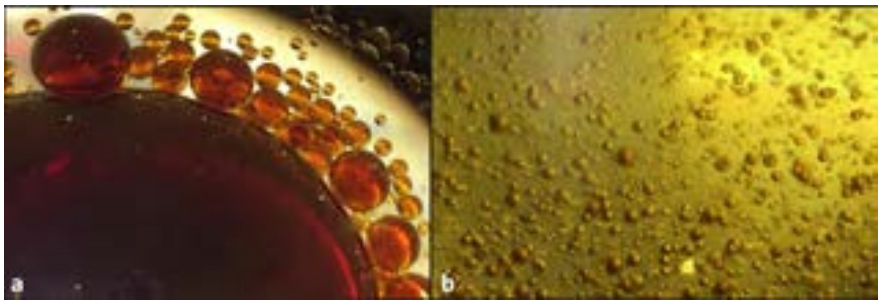

Figure 3.2: Oil (dark) and vinegar (clear) form distinct liquid phases. (a) Large droplets nestled on a surface (b) smaller drops suspended in solution.

droplet is large enough, a subregion of the droplet can then be bleached - the region is blasted with high-intensity light causing the fluorophores within this subregion to irreversible degrade. As a result, any fluorophores within this 'bleaching volume' are switched into a permanently dark state. If the droplet had material properties consistent with a solid (as opposed to a liquid) this would bleach a well-defined circle which would remain stable and dark (as an example, see Fig. 3c in Riback *et al.* [483]). However, given our vinegar droplet is liquid-like, and hence dynamic, then even though a subset of the molecules are now 'dark' they will still diffuse and exchange both within the droplet and with the bulk pool. This exchange allows the bleached region to recover from its dark state by exchanging bleached molecules for non-bleached species. The faster the recovery is, the more dynamic the droplet. A graphical overview of this entire procedure is provided by figure 3.3.

A final important idea to introduce is that of the *saturation concentration*. If we add a *tiny* amount of vinegar into our oil solution, the vinegar will disperse, and the system remains in the single phase regime, with oil and vinegar homogeneously mixed. We can slowly increase the bulk concentration of vinegar in our well mixed solution, and at some threshold concentration we will observe the formation of vinegar rich droplets (see fig. 3.4). This threshold concentration is referred to as the saturation concentration (c_s) - beyond this

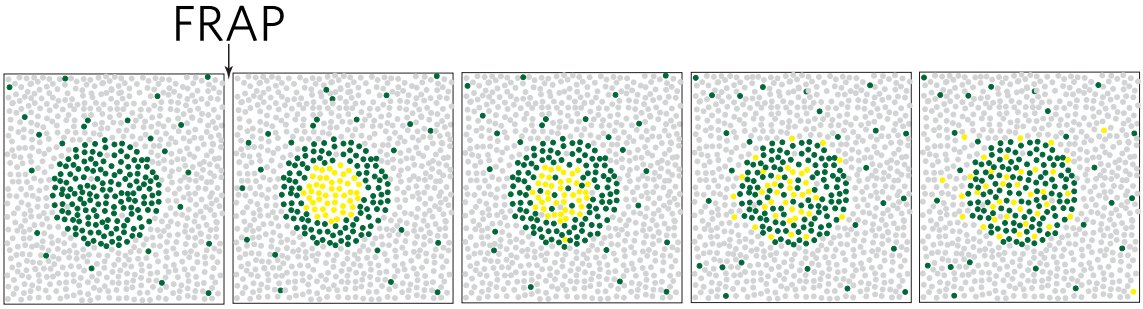

Figure 3.3: Partial Fluorescence Recovery After Photobleaching (FRAP) experiments provides an experimental approach to assess internal re-arrangement of droplets. FRAP occurs in the second frame, converting a central region of green molecules (photo-active) to yellow (photo-bleached). After the FRAP event, gradually recovery occurs due to dynamic exchange.

value, the bulk phase can no longer support additional solute. This thresholding behaviour reflects the fact that a phase transition is infinitely cooperative - you either have two phases, or you don't.

3.1.1 Demixing is Driven by Preferential Interactions

As we will discuss extensively in chapter 13, when two ideal liquids are combined the entropy of mixing is always favourable. As a result, if we combine two two liquids where neither experiences strong homotypic or heterotypic attractive or repulsive interactions they will mix to form a single homogeneous phase. To obtain a demixed system requires preferential interactions [140, 504]. Considering this, it should be clear that for our oil and vinegar example we might expect preferential oil-oil and vinegar-vinegar interactions that drive the formation of two phases. This is partly true.

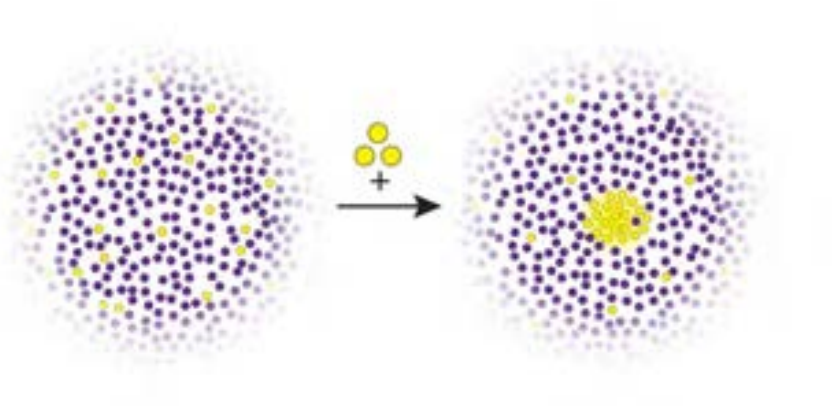

Figure 3.4: When the bulk concentration is below the saturation concentration ($c < c_s$) a single phase exists, but above the critical concentration ($c > c_s$) a separate dense phase forms (yellow region in the center)

In reality, vinegar is a mixture of acetate and water. Acetate and water share an ability to form strong hydrogen bonds, which gives rise to acetate's high solubility in water; water-water, water-acetate, and acetate-acetate interactions are all approximately equal in strength[16], such that when we mix water and acetic acid see a single homogeneous liquid. For simplicity, we'll assume 'oil' in this case is a homogeneous and mono-disperse mixture of some aliphatic polymer (say oleic acid). When we mix oil and vinegar the oil-acetate and oil-water interactions are much weaker than the water-water, water-acetate, or acetate-acetate interactions. The oil simply cannot compete with those strong hydrogen bonding interactions, and so is excluded from the vinegar. This mismatch of interaction strengths gives rise to our two-phase system.

It may seem like we're belabouring the point (and we are), but there is an important and somewhat counter-intuitive idea buried in here: we have not discussed the strength of the oil-oil interactions. They are weak. In fact, they could be non-existent, or even repulsive.

[16]For the sake of this discussion, let's assume this is true

Phase separation does not have to be driven by strong interactions associated with all of the components, but simply a mismatch between different phases. To take anthropomorphism to a dangerous level, its not that oil "likes to interact with itself", but that vinegar "doesn't like to interact with oil". This is enough. Oil might be equally happy interacting with itself or with vinegar, but if the vinegar really likes to interact with itself this will exclude the oil and drive demixing.

Does this matter? We believe so. As an example, this mismatch in interaction strengths can be used to quantitatively explain the internal architecture of the nucleolus [172]. This idea also explains why despite being considered a hydrophilic amino acid, polyglutamine is strongly aggregation prone [116]. More generally, it outlines an important idea that phase behaviour is (to a certain extent) a balancing act between relative interaction strengths with respect to the other components in the system and with respect to thermal fluctuations. In principle, this provides a mechanism through which seemingly passive players could be driven into or out of condensates for a variety of functions. If interactions are strong enough phase separation may lead to the formation of a solid or a glass, which may be desirable, or may be deleterious [52, 307, 369, 399, 406, 443, 483]. If interactions are too weak enormous concentrations of solute will be required to cross the critical concentration and form assemblies. In short, the relative strength of interactions are the *tunable* determinants of phase separation.

An important tenet associated with the physics of phase separation is that interactions that drive phase separation must be multivalent. That is, each monomer must be able to bind to more than one other partner. It is tempting to think of these interactions as weak (with respect to thermal fluctuations), and indeed for liquid-like condensates this is a requirement, but for solid-like assemblies (and even reversible amyloids, as in A-bodies) these interactions

are necessarily strong [15, 251]. It is also tempting to assume that the components involved must engage in 'highly multivalent' interactions, but in the case of work by Li *et al.* robust phase separation is achieved with low degrees of multivalency [330]. Finally, it often stated that these interactions must be non-specific, and indeed that may be true in other cases, but highly specific interactions are also perfectly able to drive liquid formation [330]. Taken together, this presents an enormous evolutionary space in terms of how condensates form, and what their material properties will be.

3.1.2 Phase Diagrams Provide a Powerful Quantitative Framework

Putting this all together, we can use phase diagrams to present a unified description of the phase behaviour of some solute (e.g. vinegar in oil). Figure 3.5 provides a simplified schematic of such a phase diagram [504].

Here, the abscissa (X-axis) reports on the bulk (total) concentration of vinegar and the ordinate (Y-axis) reports on the interaction strength between 'vinegar' molecules. This phase diagram is a two-dimensional map that reports on the phase behaviour for a given vinegar concentration:vinegar interaction strength tuple. The black curve represents the coexistence curve (also called the binodal) between the one phase and two-phase regimes, and for a given interaction strength the left-hand side of the curve defines the saturation concentration described above.

In chapters 12 and 13 we provide an expansive discussion on how to understand, use, and construct phase diagrams, so in the interest of efficiency we will not delve into those topics here. Suffice to say, phase diagrams of binary systems predict and describe an infinitely

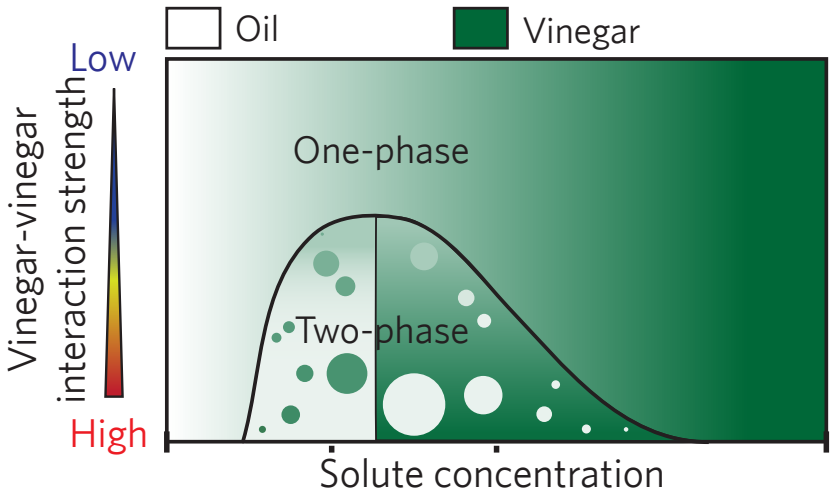

Figure 3.5: Representative phase diagram for a binary solution. The black line corresponds to the coexistence curve (binodal). The crossover point (at which the dense phase becomes the major phase and the dilute phase becomes the minor phase) is drawn here as a vertical line from the critical point, though this need not be the case.

cooperative transition between a disperse one-phase system and a demixed two-phase system. The concentration at which this transition occurs depends on the relative interaction strengths between the various components in the system. The schematic in fig. 3.5 is an idealized phase diagram for a binary system; for tertiary systems (and beyond) more exotic phase behaviour can occur, which goes beyond the scope of this introduction [256, 257].

The remainder of this introduction will provide a more qualitative and biologically focused introduction to phase separation. We will return to the associated physics in chapters 12 and especially 13.

3.2 Phase Separation in Biology

We have so far managed to introduce the physical concept of phase separation in the context of salad dressing. While convenient in terms of providing a macroscopic anchor for our introductory discussion, cells do not - as far as we know - use salad dressing to form phase separated states. Instead, biological phase separation is driven by proteins, frequently in conjunction with RNA.

3.2.1 Phase Separation and Gelation

In the interest of semantics, we will use a blanket term - biomolecular condensates (or just condensates) - to refer to these assemblies of interest. Condensates deliberately does not distinguish between gels, liquids, solids, or glasses, but instead simply implies a non-stoichiometric assembly driven by multivalent interactions. Such an assembly could also be referred to as a quinary assembly, but we will use the term condensates here to capture the essence of an inherently three-dimensional and somewhat disordered coalescence. In almost all of the cases examined thus far, we believe the initial formation of a condensate is driven by an initial phase separation or phase transition process. Formally, this may be best described as a condensation (gas to liquid), deposition (gas to solid), crystallization, (liquid to solid) or a true liquid-liquid phase separation. By gas, liquid, and solid we refer to the material state of the solutes of interest in the context of the cell; clearly we are not suggesting that gaseous proteins exist in the cell, but the physics associated with a gas (low density, weakly interacting solute) to liquid (higher denser, more strongly interacting solute) may be more appropriate than a liquid-liquid phase separation.

To some extent, the language used depends on the degree of conceptual coarse-graining the reader is willing to endure. If the cytoplasm is considered a simple liquid, and the formation of condensates now represents a second and distinct type of liquid co-existing with the cytoplasm then this can be considered liquid-liquid phase separation. If the cytoplasm is be considered a complex gas with many 'gas molecules' (proteins and RNA) diffusing around, and the formation of a condensate is truly the emergence of a liquid phase whereby one of those 'gasses' condenses, then this process would indeed be better described as a condensation. Different individuals will have different preferences for how they wish to describe these processes, and we make no judgement regarding the 'correct' verbiage to use.

The definition of a gel has a (genuinely) surprisingly convoluted history. Various definitions exist based on rheological behaviour, and while they appear convenient from an industrial perspective, from a thermodynamic perspective they are somewhat unsatisfying. We suspect a large part of the confusion surrounding the definition comes from the fact that gels appear in many different fields, from material science to polymer chemistry to dental science. For our convenience, we propose first a formal thermodynamic definition, which is consistent with Flory's work, followed by a description of what people often consider gels to be [178]. In this discussion we consider only physical gels (in which the interactions between individual components are non-covalent), although recognize that chemical gels (in which interactions between individual components are covalent cross-links) can be thought of in much the same way.

We consider phase separation and gelation to reflect physical processes that define two distinct transitions. Phase separation is a *density transition*; as a phase boundary is crossed, the density of species undergoes a discontinuous change from the dilute phase to dense phase. As mentioned, in phase separation, there is a *saturation threshold* (c_s) that defines the point

at which phase separation occurs. This definition of phase separation says nothing about the dynamics of either phase; the dense phase can be liquid or solid. In a similar vein, gelation is a *topology transition*; upon gelation, the density of our species of interest remains the same, but the underlying topology of the system undergoes a transition such that we now have a system-spanning netwok. A system-spanning network reflects the fact that there exists at least one networked structure within the gel that is of the size of the entire system. Analogous to the saturation concentration, there is a *percolation threshold*, but unlike phase separation, the percolation threshold demarks a continuous transition in topology as a function of concentration. Again, like phase separation, our definition of gelation does not make any assumptions about the internal dynamics; gels can be liquid or solid.

We extend this definition slightly based on the work of Almda *et al.* to suggest that as well as providing a connected topology with system-spanning networks, a gel should internally support a second liquid phase [6, 417]. Based on this definition, pure water would not be considered a gel (no second component), but a PEG solution above the overlap concentration would be (the PEG network supports an internal water phase). This definition of a gel is rigorous and unambiguous, although it is also broad. Based on this definition, it would seem that phase separation cannot occur without gelation[17], while gelation can occur without phase separation (as in our PEG example). With this in mind, our definition introduces an inherent coupling between phase separation and gelation. The density transition associated with phase separation leads to the formation of a dense phase, and this dense phase is likely now above the percolation threshold, meaning the dense phase is also a gel. This coupling is explored in detail in work by Harmon *et. al* [223].

[17]This has not been rigorously explored and remains an open question

While we believe this definition to be complete, in many examples gels are taken to be solid [6, 417]. This poses a problem for a thermodynamic definition - one could (rightfully) argue that to demand physical gels are solid is to argue that non-covalent bonds must be infinitely stable, a requirement that would invalidate the classification of many apparent gels [178]. With this in mind, we suggest that such a thermodynamic definition is not possible, but that a subjective phenomenological definition from the perspective of biological phase separation can be made if domain-specific information is included. In the interest of clarity, this should not define a *gel*, but perhaps an *apparent solid*. For our definition of an apparent solid, we must specify a characteristic time-period over which we require solid behaviour to be observed. The rearrangement of internal topology should be significantly slower than this characteristic time-period. For our discussion on biological phase separation we will consider this characteristic time period to be *ca.* an hour. We choose this time-scale simply to place the solid-like behaviour on biologically relevant footing; an hour is (very) roughly 10× the time it takes for a gene to be transcribed or 100× the time it takes a protein to be translated [533]. If internal rearrangement occurs on timescales shorter than this, then and we have a two-phase liquid where one of the liquids is highly viscous. If it occurs on timescales longer than this, then we have an apparent solid.

Conveniently, the molecular re-arrangement of condensates can be probed via FRAP, providing, at least in principle, a method to assess condensates within this functional framework. Note that this characteristic timescale is a lower bound, but not an upper bound - i.e. we expect there are many examples of biological gels that are solid and stable for weeks or even years [52]. Taken together, we suggest these two definitions provide accurate and rigorous definitions of a gel and of an apparent solid, and allow us to move away from the tenuous equivalence of gels as solids.

The discussions above focus on a thermodynamic description of phase separation and gelation. An addition consideration is of the kinetics of phase separation. There are various examples of spherical droplets forming that show apparent solid like behaviour, a result that is often ascribed to an initial phase separation followed by gelation [217, 282, 307, 483]. This is possible, and would suggest that phase separation can occur without gelation, at least in a non-equilibrium process (i.e. phase separation precedes gelation). However, it is also possible that phase separation and gelation are necessarily coupled, and the acquisition of apparent solid-like behaviour originates from a slow internal organization of the gel to minimize the free energy of mixing and/or distinct conformational re-arrangements between interacting species. Macroscopically, both these scenarios would be manifest as the formation of spherical droplets (phase separation) that subsequently become 'sticky', are no longer able to fuse or show rapid internal dynamics (due to internal coarsening). Importantly, when the protein bulk concentration is reduced below the critical concentration, this assembly *may* remains stable for an arbitrary period of time if the solid-like state is metastable. If this metastability is fully robust to the majority of ambient fluctuations then dis-assembly may rely on an entirely separate set of physical processes to those that were involved in phase separation. This description is fully consistent with the behaviour observed for stress granules [483, 639].

As a final confounding factor, a single protein may have multiple regions that engage in entirely distinct self-assembly processes that become coupled by the architecture of the protein. As an example, one could imagine a protein with two domains (grey and yellow in fig. 3.6). The saturation concentration for the yellow domain is substantially lower than for the grey domain, such that once a *protein* concentration above the yellow domain's saturation concentration is reached phase separation is achieved solely via the yellow domains. This gives rise to a condensate which now has a high concentration of yellow *and* grey domains.

Within this condensate, the concentration of grey domains is now above the gel point, leading to the condensate undergoing a change in network topology without a change in density and coalescing into a solid - i.e. undergoing gelation. This gel may now be stable even when the bulk concentration of protein drops below the saturation concentration of the yellow domain.

This may seem convoluted, but it is precisely what appears to be happening in the case of Pab1 in response to heat stress [483]. Moreover, from a functional perspective, this allows a tunable decoupling between the protein concentration and the assembly state, effectively providing switch-like behaviour which on the timescale of the cell could be irreversible without some active process to disassemble the gel state. Elegant work by Roberts & Harmon *et al.* (unpublished) provides a complete molecular description of how such a process could happen via a synthetic peptide system, with important implications for both materials sciences and biology.

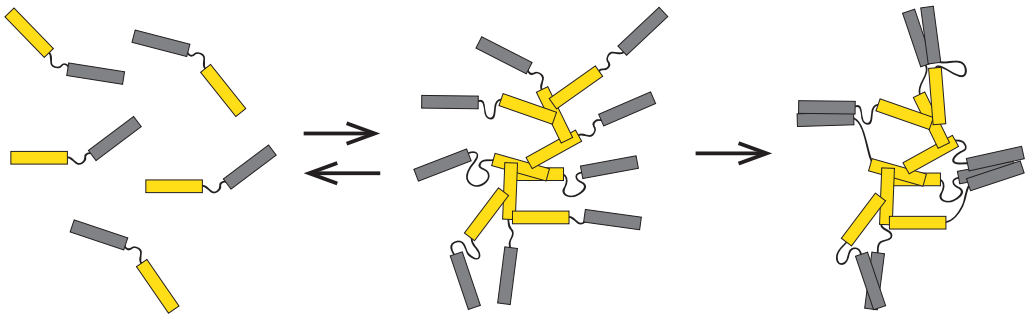

Figure 3.6: A putative two-stage model for the coupling of phase separation and gelation. Initial assembly is triggered by the yellow domains, which leads to a dynamic assembly with a local concentration of grey domains above a saturation concentration. Subsequently disassembly, is kinetically retarded by the highly networked assembly.

3.3 Biomacromolecules of Phase Separation

The first report of intracellular liquid like assemblies was by Brangwynne *et al.* in 2009 [65]. In what is already seminal work, Brangwynne demonstrated that the spatial organization of P-granules in the early *C. elegans* embryo was driven not by intra-cellular cytoplasmic flow, as had previously been proposed, but instead due to a gradient in the P-granule saturation concentration as a function of embryo position. Upon symmetry, breaking the saturation concentration in the posterior half of the single-cell embryo decreases, such that the concentration of P-granule components become supersaturated. This, in turn, leads to the condensation and formation of P-granules in a well defined and spatially regulated fashion. When shear stress was applied to large nuclear-associated P-granules the organelles dripped, flowed, fused, deformed, and showed complete internal re-arrangement on the order of seconds, characteristic behaviour of a fluid. Towards the end of the paper, the authors remark

> *We propose that P granule localization exemplifies a general mechanism for organizing the cytoplasm that arises from collections of weakly 'sticky' molecules, including other ribonucleoprotein assemblies (e.g., P bodies, Cajal bodies, or stress granules).*

In the following eight years it has become clear that the formation of intracellular condensates through phase separation is an abundant and ubiquitous process in biology.

While this work represented the first demonstration that intracellular assemblies exhibiting liquid-like behaviour, the formation of dynamic, disordered assemblies as functional entities in biology was proposed as early as the 1950s [54]. Even before that, in an alarmingly prescient article published at turn of the 19th century, Wilson suggested that the protoplasm

(cytoplasm) would be a mixture of liquids [645]. Similarly, the ability of proteins to undergo liquid-liquid phase separation is not a new phenomenon and has been well known by the crystallographic community for decades [69]. Indeed, as recently as 2005, dynamic assemblies that show all the hallmarks of phase-separated liquids were characterized in the context of Wnt signalling [529]. We have not performed a rigorous search through the literature for other such examples, but it seems likely that micron-scale liquid-like assemblies have been repeatedly discovered, but without the coupling of a theoretical framework the functional implications of these large, dynamic assemblies likely seemed obscure.

Many previously identified membrane-less organelles including the nucleolus, paraspeckles, nuclear speckles, cajal bodies, PML bodies, P bodies, stress granules, and germ granules - as well as various other cellular assemblies - have now been shown to display features consistent with condensates [75, 106, 172, 184, 314, 315, 421]. Several of the more well characterized examples having been intricately explored and shown to posses partial or complete liquid-like behaviour [65, 172, 421, 639]. Like P granules, these condensates are large (typically around 1-2 µm in diameter) and compositionally heterogeneous, composed of many different components [8,261,601,622]. For these larger organelles, RNA is also known to be important, with several organelle-specific RNAs required for function and assembly [109, 668]. Is RNA a necessary component of large, micron-scale intracellular biological condensates? In recent work discussed in 11 we demonstrate that, at least in principle, intracellular organelles do not require RNA to form. Moreover, extensive *in vitro* studies on many proteins have clearly demonstrated for a wide range of systems a single protein (or even a single domains) is both necessary and sufficient to drive phase separation [51, 79, 162, 338, 399, 421, 443].

There are numerous examples of proteins identified by chance, through screens, or based on mechanistic hypotheses that have been shown to form liquid droplets *in vitro* [51,79,217,266,

282, 338, 369, 399, 406, 443]. For many of these proteins the relationship between phase separation and the protein's normal functional role remains unclear, but in a number of cases the identified proteins are known to be associated with inclusion bodies in histological samples taken from patients with neurodegenerative diseases. The link between phase separation and diseases remains highly correlative, but minimally causative. While these proteins have been implicated as causal factors, it remains entirely unclear if their aggregation simply reflects a cellular-wide shut-down in the proteostatic machinery [340].

There has been an inherent focus on large condensates associated with the study of phase separation *in vivo*, driven in no small part by convenience - big things are easy to visualize by light microscopy, allowing direct visual characterization. Recent developments in super-resolution technology have begun to shed light on condensates that exist below the diffraction limit [107]. While not yet fully characterized as liquid-like, cluster of RNA POL II have many of the traits expected for assemblies driven by phase separation [100]. More generally, we anticipate that there may be a wealth of cellular condensates that form via protein-only phase separation, yet are below the size resolution accessible to conventional microscopy. Finally, recent work from Jain and Vale suggests that repeat-length dependent RNA-only phase separation leads to the formation of apparent solid condensates *in vitro* and in *in vivo*, adding an entirely new role for RNA in phase separation [260]. Various aspects of DNA biophysics have been examined in the context of phase separation. Notably the protein Ki-67 has been shown to act as a molecular surfactant to aid in maintaining chromosomal solubility [119]. The synaptonemal complex - the cellular apparatus involved in the exchange of genetic material between chromosomes in meiosis - has been show to assemble and disassemble in a manner analogous to a liquid-crystal [493]. We quietly anticipate a barrage of papers exploring chromatin organization and transcriptional regulation through the lens of phase separation [232]. Indeed, work on chromosomal territories and the non-equilibrium

fractal globule behaviour of condensed nuclear material hints at the possibility that nuclear organization is facilitated at least in part through the formation of many globally-immiscible liquid phases.

Taken together, an emerging picture of biological condensates is one in which protein and nucleic acids work in concert. Many of these assemblies appear to show liquid properties, including rapid re-arrangement times and exchange with bulk solutes, as well as droplet fusion, wetting and dripping. However, we caution against an obsession with the material state of condensates. Nature selects for function, and function alone. While it is entirely reasonable that internal dynamics and droplet concentration may be under evolutionary pressure in some cases (see chapters 12, 13), a wide range of material states are likely to be functionally relevant [56, 172, 483].

3.4 Biological Phase Separation as a Means for Cellular Organization

Why might cells use membrane-less organelles instead of membrane-bound ones? While absolute arbitration on the answer to this question is likely not possible, we propose that the functional role of membrane-bound organelles reflects an association with biochemical processes where there is a need for chemical protection. This could refer to the protection of the interior of the organelle from the chaos and noise of the cytoplasm (nucleus, golgi, enodplasmic reticulum, mitochondria, chloroplast) or where there is a need to protect the cell from the chemistry associated with the organelle (mitochondria, chloroplast, lysozome, peroxisome). More generally, this distinction may reflect a need to fight passive diffusion.

The evolutionary path that led us to the current cellular architecture was not a well thought out exercise in process design, but the path of least resistance. Consequently, we may wish to ascribe logic to the partitioning of cellular function between various organelles, systems, and processes, but it is unclear if this is an appropriate (or even relevant) course of action.

Although largely not discussed in this work, we are not discounting the fact that in many cases the formations of biomolecular condensates may be driven by an active, energy dependent process [41, 169, 678]. The coupling between active processes (which provide intimate control) and phase separation is likely critical, and provides a means for these condensates to move far from thermodynamic equilibria. This in turn may be *the* mechanism through which these condensates can perform work. The field of active matter has many ideas and principles to offer those working on biological phase separation, and while we do not consider it further in this work, the importance of non-equilibrium statistical mechanics in understanding the functional importance for biological phase separation will likely be significant [362].

A useful terminology for thinking about these condensates was proposed by Banani *et al.*, who suggested that the components (RNA or protein) that are necessary and/or sufficient for the formation of condensates be referred to as *scaffolds*, while components that will selectively partition into condensates once they have formed are designated *clients* [28]. What role(s) might phase separation play in biology? Here we summarize various proposed and putative roles of intracellular condensates in cellular function.

3.4.1 Compartmentalization

The most obvious function for large (micron-scale) cellular condensates is that of compartmentalization. The intra-droplet environment is expected to provide a unique and

condensate-specific environment. Demonstrating the functional relevance of this postulate has been challenging; the best example to date is from Nott *et al.*, who showed that the interior of the Ddx4 droplet significantly reduces the free energy associated with double stranded DNA melting, providing an environment that is equivalent to around 4 M GdmCl [420]. Given the preferential accumulation of certain components within these droplets one putative hypothesis is they provide a mechanism to concentrate various enzymes to create highly efficient micro-reaction environments for enhanced chemistry. Such a hypothesis is not without precedent: the bacterial carboxysome encompasses a proteinaceous outer shell, and concentrates the various enzymes involved in the Calvin cycle into its interior, providing a micro-structure to facilitate catalysis in a manner akin to platinum catalysts in industrial applications [516, 571]. The functional significance of compartmentalization could be further enhanced by distinct patterns of spatial organization within the droplet, allowing for sub-compartments associated with distinct chemistry [172, 390]. In work not included in this thesis, well-defined spatial organization has been observed in nuclear speckles[18]. We expect that many complex organelles are likely to contain significant internal organization, partly through design, and partly because generating fully miscible multi-component droplets places an enormous evolutionary constraint on those components, where that constraint grows exponentially as additional components are included.

Although phase separated micro-reactors represents an appealing cellular design approach, we suggest a word of caution. With the exception of the nucleolus, true functional demonstrations of these organelles performing specific chemistry at significantly greater efficiency remains surprisingly lacking. This is likely in large part simply due to the technical challenges associated with tracking these reaction. Despite this, a curious lack of phenotype is frequently observed when these organelles are genetically ablated [162]. The phenotype

[18]Fei, ..., Holehouse, *et. al* (unpublished)

of cells under rich growth in a controlled environment is, arguably, not necessarily a useful measure of true fitness. Nevertheless, given the apparent complexity associated with these organelles it is somewhat surprising that their disruption does not have a more significant impact of basal cellular function. Unpublished (but reported) results from the Rosen group suggest that, *"the highly concentrated scaffolds and enzymes within phase-separated droplets frequently interfere with each other, with scaffold components inhibiting enzyme activities and enzymes dispersing droplets by covalently modifying scaffolds"*, although they note that *"in cells, it is likely that mechanisms exist to prevent or take advantage of such interference"* [27].

3.4.2 Sequestration

Sequestration remains one of the most well demonstrated examples of function associated with cellular condensates. Stress granules are believed to represent a complex response to stress conditions. Part of this response involves sequestering folded proteins and bulk mRNA to aid in a reprogramming of the cellular translational network away from normal function and towards the heatshock response programme [77, 261, 369, 483, 622]. Similarly, sequestration on a cellular[19] or sub-cellular[20] level appears to be a commonly used mechanism for the temporal regulation of cellular behaviour, allowing cells to 'jump' in time by robustly but reversibly depleting the cytoplasm of various components in specific or non-specific manner [15, 52, 483].

Indeed, recent elegant work by Riback, Katanski, *et al.* in *S. cerevisiae* provides one of the strongest links between condensate formation, function, and phenotype to date [483]. Their work focusses on the polyA binding protein(Pab1), which is normally soluble and binds to

[19]Boothby, ..., Holehouse, *et. al* (unpublished)
[20]Powers, Holehouse, *et. al* (unpublished)

mRNA transcripts with A-rich 5' untranslated regions (UTRs), an RNA feature strongly associated with heatshock proteins. Pab1 shows a remarkable response to heat-stress and pH stress, forming spherical assemblies with apparent solid characteristics instantaneously in response to temperatures even a few degrees above the normal growing temperature. Upon condensate formation, the previously bound heat-shock response RNAs are released, leading to a burst in the translation of heat-response transcripts. Mutations that modulate Pab1's ability to form condensates directly dictate cell fitness. This is likely only part of a complex stress response, with various other factors playing related but distinct roles in remodelling transcription, translation, and cellular status [76, 77, 134, 400, 622]. Work by Jain *et al.* suggest that stress granules contain a liquid-like periphery with a more solid core which may form via specific protein-protein and/or protein-nucleic acid interactions, as opposed to weak, non-specific interactions [261, 639]. It is entirely possible that this core acts as a nucleation point, facilitating the formation of a less specific and liquid-like shell around the core that encompasses a range of cellular proteins to protect them from the perturbed stress environment.

3.4.3 Concentration Homeostasis

A related but distinct function is that of concentration homeostasis. An inherent property of liquid-liquid phase separation is the rapidly responding and energy independent maintenance of the bulk concentrations associated with the various species that are associated with condensate. These bulk concentrations are, by definition, held at their saturation concentration. Such a mechanism at least in principle provides an attractive mechanisms for protein and RNA concentration homeostasis inside the cell; excess components are sequestered into droplets, while during times of bulk depletion can be released from droplets.

A slight wrinkle in this hypothesis is that thus far, dynamic, liquid-like intracellular condensates have not been observed in bacteria. If this were an evolutionarily powerful mechanism for cellular regulation, we would expect it to be present in prokaryotes. Of course, a possible explanation for this is simply one of scales. Bacterial are substantially smaller than eukaryotic cells; the cell volume of *E. coli* is around $1\,\mu m^3$, while the volume of *S. cerevisae* is around $42\,\mu m^3$ [272, 451]. While the cell volume is low, the proteins are the same size, which may make intracellular condensates invisible to conventional microscopy, unachievable (absolute magnitude and timescales associated with cellular concentration fluctuations are too great for droplet formation to be obtained) or unnecessary (bacterial cells are small enough and transcription and translation can be coupled such that protein synthesis is directly regulated at the gene level with minimal post-transcriptional regulation). Similarly, the timescales at which bacterial processes occur on may simply mean that condensate formation and dissolution would add a latency that provides a selective disadvantage [533]. However, we do anticipate that liquid-like condensate formation in bacteria will be identified, although it may be less prevalent than in eukaryotic cells.

3.4.4 Integration of Complexity

A hypothetical explanation for the formation of condensates is that these complex assemblies allow for an energy independent mechanism to integrate complex analog signals into a single digital output. In an abstract sense, one could consider a multi-component phase separated condensate as a complex `AND` gate, whereby all the components need to be 'on' (above a critical concentration) for condensate formation to occur. If we take our circuit analogy to its logical conclusion, the output signal could be some process or event that occurs upon

formation of the condensate, but not before. We suggest three possible reasons why, from a signalling perspective, this might be attractive

1. Unlike a true `AND` gate, it may be possible to compensate for lower levels of one component with higher levels of another building direct and energy independent redundancy into the system. For example, complex signalling pathways could be integrated in this way, such that any one of many possible inputs leads to a single output. In a similar vein, the actual function of the condensate could be tuned by the relative proportions of different components, allowing an apparently single condensate to perform different function depending on its state (where state reflects the relative proportions of different constituents).

2. Conversely, one could also envisage a system whereby one particular component is absolutely required for condensate formation. For such a system, the remaining semi-redundant components may exist at a high concentration, and upon the appearance of this key component a complex multi-component condensate can form immediately. This would provide a mechanism to drive an enormous and rapid amplification in process complexity, allowing a single input signal to drive the formation of a complex output signal in a binary manner.

3. Finally, from a signal-processing perspective, if the output is some binary event that depends solely on the presence or absence of a condensate, then phase separation provides a way to entirely suppress fluctuations in an arbitrarily large number of components to effective perform intracellular signal-attenuation. In this way, condensate formation could be regarded as a biological analog-to-digital converter. Such a behaviour would be well suited for critical decision making in cells (e.g. in differentiation, apoptosis, cell division, *etc.*).

Are there real-world examples of signal integration by biological condensates? Phase separation driven by the proteins BugZ and SPD-5 appear to play key roles in the organization and assembly of cytoskeletal components [266,649]. In the case of the protein BugZ, droplets are formed in *X. laevis* during mitosis; these droplets bind microtubules, acting in a manner analogous to the spindle apparatus in mammalian cells. The *C. elegans* protein SPD-5 appears to be the driving component in the formation of microtubule arrays which eventually assemble into centrosomes. These are both key events that require tight regulations and integration of multiple signals. We tentatively suggest these could represent phase separation acting as a master integrator of information.

RNA POL II forms clusters of around eighty molecules at transcriptional start sites; the lifetime of these clusters correlates directly with the number of mRNA transcripts produced [100]. It is hypothesized that this clustering is driven by the disordered C-terminal tail of RNA POL II. FUS, another protein known to drive liquid-liquid phase separation, is known to interact with the RNA POL II C-terminal tail based on biophysical assays [79,528]. Based on genetic screens, the loss of FUS leads to the accumulation of RNA POL II at transcriptional start sites [527]. Taken together, a model whereby FUS modulates the lifetime of RNA POL II clusters by influencing their stability emerges as a putative general mechanism for gene regulation.

In the fungi *Ashbya gossypii*, the protein Whi3 shows RNA specific clustering to facilitate multiple different cellular processes, providing an elegant example of spatio-temporal regulation by liquid-like droplets in a feedback network that relies on specific RNA [319, 668]. The work required to uncover this one example was substantial, and represents a complex interplay of protein and RNAs, but we so no reason to assume that this is the exception and not the rule.

3.4.5 Partitioning of Components During Cell Division

For soluble cellular components, equal partitioning between daughter cells during mitosis can be passive, and simply rely on the entropy of mixing to ensure that the two halves of a mother cell have approximately equal levels of the components of interest. Membrane-bound organelles, on the other hand, may need specific mechanism to ensure equal partitioning, which can be achieved through autonomous division or active partitioning [567, 632, 660]. Intracellular condensates could provide a mechanism to achieve the best of both worlds. During normal cellular function these organelles exist as well defined entities, allowing them to perform their characteristic function(s). Upon mitosis they dissolve, allowing their constituents to mix with the cytoplasm, be partitioned equally, and then re-condense in the daughter cells once mitosis is complete.

A convenient feature of such a mechanism is that as cells begin to divide cellular volume increases substantially [677]. This in turn could lead directly to dissolution of condensates due to the cellular concentration of scaffolding components dropping below the saturation concentration. In effect, this provides a 'free' mechanism to distribute complex organelles equally between two daughter cells. Freedom here reflects both the absence of an energetic cost (beyond the energy needed to expand the cell volume), and the fact that no specific (or complex) cellular apparatus is needed. Moreover, via the converse process, this provides a potentially simple mechanism for asymmetric partitioning of soluble species, as evidence in P-granules during *C. elegans* embryogenesis [65]. Demonstrating that this equipartitioning hypothesis is an evolutionarily selected for feature (and not just a necessary correlate) would be challenging. However, given that the converse has already been observed in the case of P-granules, we might expect other examples of droplets forming asymmetrically across a division axis to exist.

3.4.6 The Default Behaviour

Despite the various functional explanations provided above, one model that we have not seen proposed - and is a much simpler explanation - is simply that the formation of condensates is a necessary default behaviour in a highly concentrated cytoplasm. The rational behind this model is that, far from a behaviour that is selected for and maintained for functional purposes, the formation of intracelullar condensates is simply an emergent property of having a highly concentrated cytoplasm, effectively leading to a cellular saturation threshold being reached. This model is largely inconsistent with apparent evolutionary selection for regions that are necessary and sufficient to drive condensate formation. We believe it is inconceivable that this a *universal* explanation, especially as we enter into a new stage in which functional roles are being increasingly identified for these condensates.

Nevertheless, it remains a possibility that at least a subset of these condensates are simply a consequence of a highly concentrated cytoplasm, leading to the precipitation of protein and/or RNA condensates. If this were the case, it is conceivable that additional proteins (such as those with low complexity domains) may have evolved to help maintain the reversibility of these condensates, acting as non-specific chaperones and using their polar-rich IDRs as highly targetable local denaturants (Q/N/G are chemically similar to urea, R to GdmCl). This would give rise to an apparent evolutionary selection for condensate-forming proteins, but in reality the converse occurred; the formation of non-specific protein condensates gave rise to the evolution of these low-complexity proteins to facilitate condensate dissolution. Similarly, the over-abundance of RNA binding and helicase domains in the proteins associated with condensates could reflect their role as large non-specific RNA chaperones. Recent work by Jain *et al.* demonstrate that RNA alone can form condensates with solid-like characteristics [260]. Given the concentration of RNA in the cell, it seems

entirely plausible that many of these assemblies are maintaining RNA-dynamics in an ATP dependent manner. We reiterate that this is complete speculation, and cannot account for the range of examples where well defined functional behaviour associated with condensates has been identified and discussed in this chapter. However, as mentioned, the evolutionary trajectory of complex multicellular organisms is not based on design, but on relative fitness.

3.5 Sequence Determinants of Protein-Mediated Phase Separation

What types of proteins drive condensate formation. In many (though not all) of the proteins identified as scaffolds or shown to form condensates *in vitro*, large intrinsically disordered regions are necessary and frequently sufficient to drive phase separation [162, 217, 282, 399, 406, 443]. Disorder provides a convenient structural state to mediate the kinds of weak, multivalent interactions that are expected to be important for the formation of liquid-like condensates. Similarly, it offers an opportunity to create distinct sticky-patches along a chain, giving rise to a biological manifestation of the 'stickers on a chain' concept, proposed by Semenov and Rubinstein in the late 90s (fig.3.7) [505, 532]. We will return to this conceptual framework extensively in chapters 12 and 13.

While disordered regions are frequently associated with the formation of cellular condensates, they are not required [330, 483]. Surprisingly there are many published and unpublished examples of folded domains either driving the formation of cellular condensates or acting as obligate partners in condensate formation [330, 483, 668]. This suggests that the synergy between folded domains and disordered domains plays a key role in determining the phase

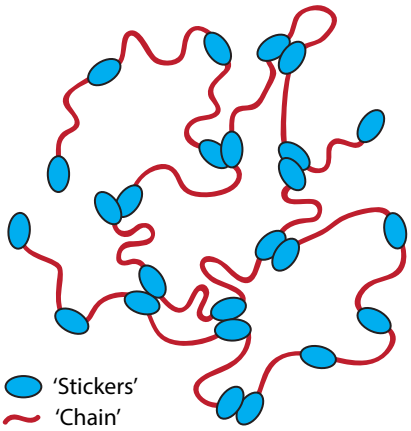

Figure 3.7: Visual representation of Semenov and Rubinstein's stickers on a chain. The linkers are flexible and non-attractive (although they could be repulsive) while the stickers engage in inter-molecular and intra-molecular interactions. Stickers could be uniformly 'sticky' for one another, or show sticker specificity (homotypic or heterotypic)

.

behaviour and material state of these condensates, and represents an entire set of open questions.

There are various different interaction types that are known to play a role in facilitating phase separation. These are highlighted in fig. 3.8 and discussed below.

3.5.1 Electrostatics

IDPs are typically enriched in charged residues. As a result, charged interactions feature prominently in several proteins (folded domains and IDPs) that have been shown to be necessary and sufficient for phase separation [162, 421, 436, 483]. One possible reason for this

is that electrostatic interactions can be modulated through post-translational modification. Recent work by Aumiller showed that reversible droplet formation between protein and RNA was possible *in vitro* using phosphorylation and dephosphorylation to trigger and then reverse protein:RNA complex coacervation (complex coacervation is discussed at length in chapter 11) [16]. An extensive body of work from the McKnight lab has strongly implicated phosphorylation as a mechanism for the reversible control of association with condensates [312, 313]. While more work is required, it seems likely that phosphorylation will play a major role in the dynamic assembly and disassembly of condensates.

The patterning of charged residues also appears important for determining the driving force for phase separation. The disordered N domain of Ddx4 is necessary and sufficient for phase separation [421]. The driving force for phase separation can be ablated by changing how charged residues are distributed along Ddx4 sequence. This provides a clear link between the single-chain and collective chain behaviour that has recently been explored further by Lin *et al.* [126, 335]. In chapter 11 we demonstrate that more than simply inhibiting the driving forces for phase separation, we are able to tune this driving force up or down by modulating the distribution of charged residues, suggesting complete control of the phase diagram.

3.5.2 Cation-pi and pi-pi

Cation-pi (specifically Arg-Phe and Arg-Tyr) and pi-pi interactions have been specifically implicated in a number of proteins. Most clearly, again in the disordered N-terminal domain of Ddx4, a phenylalanine to alanine mutant was unable to undergo phase separation [421]. In unpublished data from several groups cation-pi and pi-pi interactions are strongly implicated in a range of systems that drive phase separation. In chapter 11 we performed an extensive

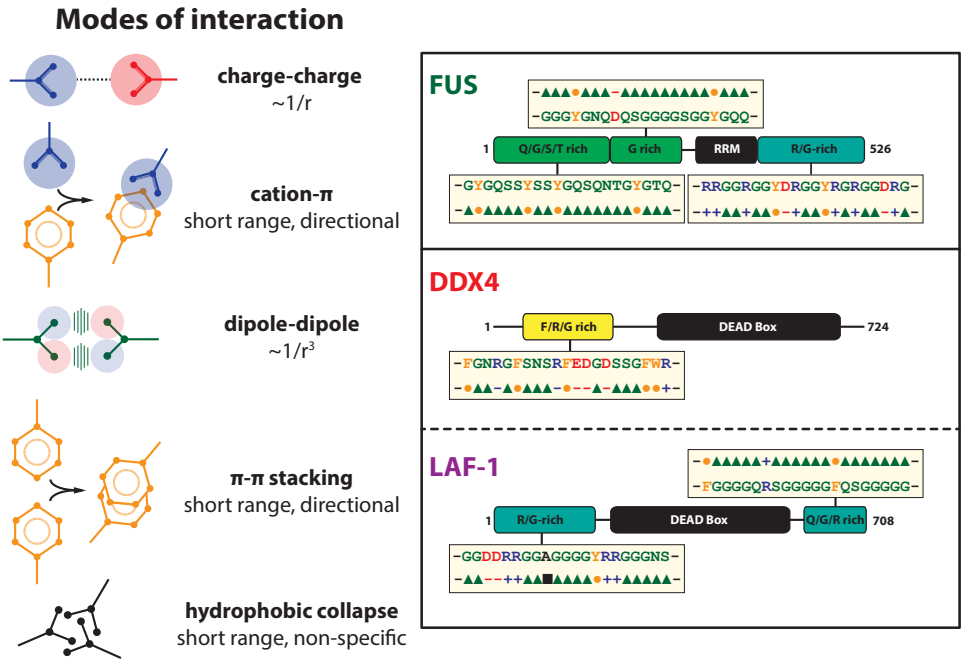

Figure 3.8: Summary of the key modes of interaction believed to be important in condensate formation. Adapted from [67][22].

molecular dissection of the residue types that were believed to be critical in mediating the formation of NICD droplets. Although NICD has a net negative charge and phase separates via complex coacervation with positively charged counterions, we unequivocally found tyrosine and arginine to be, at a residue level, the strongest influencers on the ability to phase separate. In work on synthetic polymers cation-pi interactions would found to be strong enough to overcome like charge repulsion, although we speculate a complexing counterion such as phosphate may also be involved [289].

3.5.3 Polar Interaction

Polar interactions are believed to be critical in a range of IDPs that drive phase separation. Notably, the low complexity domains of FUS and hnRNPA1 are necessary and sufficient to drive phase separation, albeit at relatively high protein concentration and are almost entirely devoid of charged residues [79, 217, 282, 399, 406, 443]. While pi-pi interactions are likely to be important, polar interactions between sidechain and backbone amide groups are also expected to play a significant role. The exact nature of these polar interactions, their relative strength, and the impact of sequence patterning remain open and important questions that can best be addressed through extensive sequence design and sequence mutations.

The relationship between these polar residues (and tyrosine) and short β sheet formation is also unclear. Recent work from the Eisenberg lab has suggested that Low Complexity Amyloid-like Reversible Kink Segments (LARKS) may allow the transient formation of β-like structure, providing a plausible explanation for the relationship between β-sheet formation and phase separation. These interactions are distinct from the structural assemblies associated with long amyloid-like polymers as observed in hydrogels, whereby multiple proteins are predicted to form fully or transiently stable linear polymers [217, 282]. In contrast, LARKS provide a rational connecting the macroscopic organization of hydrogels and the local interactions that allow for the formation of dynamic liquid-like assemblies under physiological conditions [79, 399, 406, 443]. Of particular interest, they offer a mechanism to encode distinct motif-motif specificity, in terms of chiral and directional interactions. They also provide a putative explanation for the impact of single point mutations on condensate behaviour [399, 443]. However, in their current incarnation, LARKS are considered only to mediate homotypic interactions, a constraint we do not fully understand, but one that seems arbitrary and potentially misleading [248].

3.5.4 Hydrophobic Interactions

While hydrophobic interactions are generally not associated with IDPs, in our work on NICD we found a significant correlation between the loss of hydrophobic residues and a reduction in the ability to phase separate. More recently, hydrophobicity has been shown to modulate the heat response of Pab1 [483]. Interestingly, while the specific amino acid sequence associated with the disordered P-domain is poorly conserved across different organisms, the extent of hydrophobicity is maintained, albeit through a variety of different hydrophobic residues. In elastin-like polypeptides, modulation of hydrophobicity and charge allow for the tuning of LCST and UCST behaviour [469]. Taken together, although IDPs tend not to be enriched in hydrophobic residues, these results suggest that hydrophobicity could be used to tune the temperature dependence of IDP phase behaviour, as well as drive the formation of denser droplets through a macroscopic hydrophobic effect.

3.6 Final remarks

We end this chapter with a brief summary of the general ideas discussed. We introduced phase separation, initially as liquid-liquid phase separation, and then more generally in terms of the various types of condensates that could form. We are deliberately avoiding a discussion of the material states of these droplets. In many cases there is strong evidence that condensates are indeed liquid-like, but for others the distinction between a a solid and a highly viscous liquid are challenging to ascertain. We then introduce several putative functions for phase separation in biology, and end by discussing the molecular driving forces associated with phase separation.

Part II

Single Chain Behaviour

Chapter 4

Resources to Obtain, Analyze and Classify IDPs

The following section is taken from the paper **CIDER: Resources to Analyze Sequence-Ensemble Relationships of Intrinsically Disordered Proteins** by A.S. Holehouse, R.K. Das, J.N. Ahad, M.O.G. Richardson, and R.V. Pappu. This was published in the *Biophysical Journal*, Vol. 112, pages 16 - 21, in January 2017. The text has been expanded to include additional detail. Parts of early versions of the code were developed by R.K.D, J.N.A. and M.O.G.R. All other components were performed by A.S.H.

4.1 Background

Intrinsically disordered proteins and regions (collectively referred to as IDPs hereafter) make up approximately 30% of eukaryotic proteomes [437]. They arcae associated with a variety of functions including transcriptional regulation, cell signaling, chaperone activity, regulation of bacterial homeostasis and lifecycles, viral infectivity, and subcellular organization in

eukaryotic cells [304,391,654]. IDPs are also associated with a wide range of diseases including neurodegeneration and cancer [605]. Sequence-encoded conformational heterogeneity is a defining feature of IDPs. Properties of conformational ensembles are quantified in terms of average sizes, shapes, local secondary structural preferences, patterns of inter-residue distances, and amplitudes of conformational fluctuations. Heuristics extracted from biophysical studies can be used to classify sequence-ensemble relationships of IDPs. These relationships are governed by the amino acid compositions and sequence patterns within IDPs. Recent studies have shown that sequence-ensemble relationships of IDPs contribute directly to their biological functions [127, 608].

IDP sequences show poor conservation across orthologs [94]. However, there is growing evidence that coarse-grained sequence features are well conserved in IDPs. These coarse-grained sequence properties, which can be readily deduced through analysis of primary sequences, determine the conformational properties of IDPs / IDRs. The precise sequence-to-ensemble relationship is governed by their amino acid compositions and sequence patterns [127]. Sequences encode the patterns of long range interactions, secondary structural preferences, and fluctuations about well-defined conformational elements that characterize IDP ensembles [57, 513]. Accordingly, the ensembles of many IDPs can be partitioned into distinct conformational classes, and the relationships between sequence and conformational classes can be identified using a set of quantitative heuristics that are derived from amino acid sequences [127]. The volume of sequence information is growing exponentially and hence it should be possible to uncover the evolution of sequence-ensemble-function relationships across disordered proteomes.

Low overall hydrophobicity is a defining feature of many IDP sequences. In a two-parameter space defined by the mean hydrophobicity (H) and mean net charge (q) Uversky *et al.*

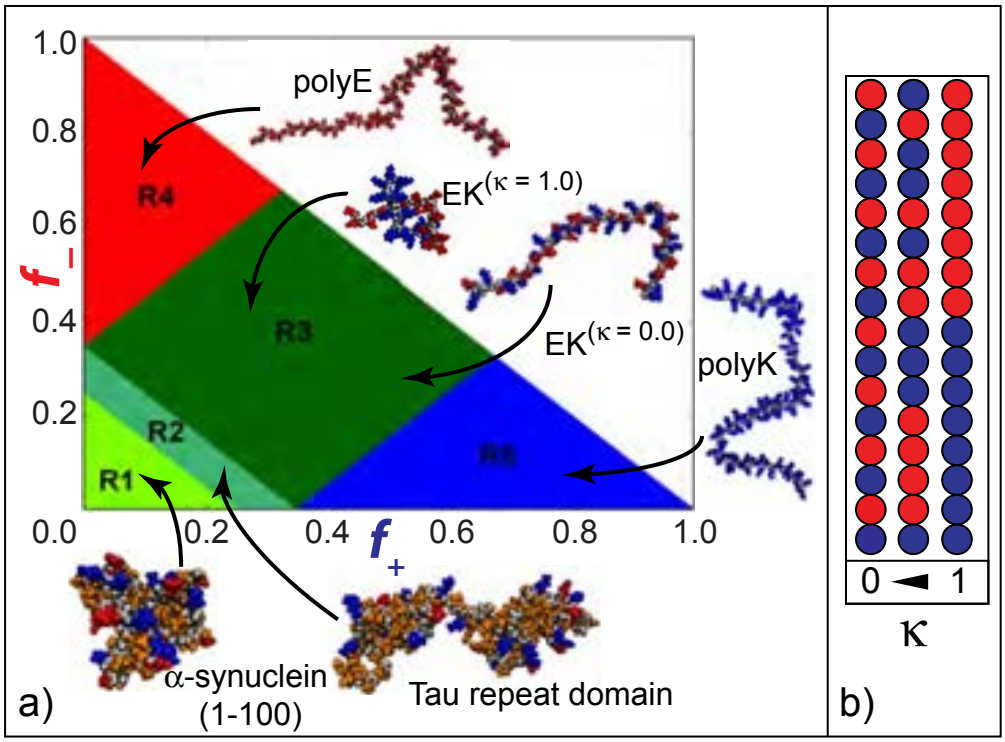

Figure 4.1: Panel (a) shows the diagram-of-states annotated with representative conformations for specific IDPs that correspond to each of the five regions. (b) shows a schematic depiction of the implication of changing κ values. Here, red and blue circles represent negatively charged residues and positively charged residues, respectively.

argued that a single empirical line delineates putative IDPs and autonomously foldable proteins [603]. Studies focused on sequences that lie on the IDP side of this empirical line showed that there are distinct sub-classes amongst IDPs themselves. For example, the net charge per residue (NCPR) of an IDP contributes directly as a determinant of overall global dimensions [359, 364, 405]. We originally suggested that polyelectrolytic IDPs with NCPR below a threshold value of 0.25 adopt compact globular ensembles, whereas sequences that lie above this threshold adopt well solvated expanded coils and even stiff rod-like conformations. However, more recent results suggest that this NCPR threshold of 0.25 is far from a

fixed rule, with IDPs with very few charged residues showing expanded behaviour [189,366]. These results suggest that the sequence determinants of collapse are more complex than previously suggested, although charged residues still play a critical role. The degree of conformational heterogeneity within IDP ensembles can be decoupled from the overall size, shape, and local conformational preferences. For example, sequences that predominantly favour collapsed globules can sample vastly different globular conformations and have higher conformational heterogeneity than highly charged polyelectrolytes that sample predominantly rod-like conformations [348]. The importance of charged residues as one of the main determinants of conformational properties of IDPs was further underscored in work that showed that the fraction of charged residues (FCR) and the linear patterning of positively charged and negatively charged residues contribute directly to the size, shape, and amplitudes of conformational fluctuations of polyampholytic IDPs [126].

Using the fraction of positively charged and negatively charged residues - f^+ and f^-, respectively - IDP sequences can be partitioned into one of five different conformational classes. This predictive, albeit heuristic diagram-of-states, shown in fig. 4.1a, provides a simple way to classify IDPs and generate expectations regarding conformational properties [126, 127]. Assuming fixed charge states, IDP sequences of low overall hydrophobicity and low overall proline content (less than 15%) can be partitioned into one of five classes: R1 - R5. Additionally, the sequence patterning of oppositely charged residues contributes directly to the global compaction or expansion of IDPs, and this patterning is quantified by a parameter κ [126]. Here, $0 \leq \kappa \geq 1$; low values of κ correspond to sequences - for a fixed amino acid composition - wherein the oppositely charged residues are well-mixed within the linear sequence. In contrast, large values of κ correspond to sequences where the oppositely charged residues are segregated into blocks of like charge see fig. 4.1b. In addition to charged residues, the fraction of proline residues and the intrinsic propensities of individual residues to adopt

polyproline II (PPII) conformations are thought to play an important role in driving local conformational transitions and global compaction / expansion [364, 582]. Finally, recent studies have focused on the sequence complexity of IDPs due to growing interest in divers of the formation of biomolecular condensates and membraneless organelles [67, 172, 391].

Our goal is to enable efficient annotation of various sequence features of IDPs and to facilitate the rapid design of sequences that enable direct investigation of sequence-ensemble-function relationships. Accordingly, we have introduced a pair of tools to annotate IDP sequences by their expected sequence-ensemble relationships. CIDER, which stands for Classification of Intrinsically Disordered Ensemble Relationships, is a web server that provides instantaneous access to a range of properties that are derivable from the primary sequence of IDPs. This includes NCPR, FCR, κ values, hydrophobicity, compositional bias, and diagram-of-states classification. localCIDER is a locally installable software package for the high-throughput analysis of disordered sequences, and includes a wider range of IDP-specific sequence analysis routines.

4.1.1 Automated Sequence and Metadata Retrieval Tools

Beyond analysing amino acid sequences, simply obtaining sequences in an automated fashion can be a major challenge. In the last ten years, large-scale analysis of proteomic data and metadata has evolved from a highly specialized research endeavour performed exclusively by the bioinformatics community into a standard component of many analyses in modern research. This change has been brought about through the combination of advances in computational power, better network (Internet) connectivity, and an explosion in the number of protein records in freely accessible databases. As an example, between 2010 and 2014, the

number of records in the UniProt database grew from ten million to approximately eighty million [600].

As publicly accessible databases have grown, they have played an increasingly important role in providing large-scale biological context for a wide variety of questions. In parallel, there has been a general shift towards the adoption of network-based application programming interfaces (APIs) which operate under the principle of Representational State Transfer (REST). RESTful APIs allow database providers to construct a single, high performance hypertext transfer protocol (HTTP) based network interface, which can be queried in a programming language agnostic manner. In this way, users can develop analysis pipelines that, via RESTful interfaces, have direct access to complete datasets using network-based API calls.

To meet the growing availability of data we designed, developed and released a general purpose library for interfacing with the UniProt and NCBI databases. Geeneus is simple to use, abstracts all of the networking and data-processing, and provides an interactive library for obtaining sequence information in an automated manner.

4.2 Methods

4.2.1 CIDER

CIDER is a user friendly, modern web server that enables rapid analysis of IDP sequences to generate expectations based on prior observations regarding sequence-ensemble relationships. It is freely accessible via http://pappulab.wustl.edu/CIDER. Full documentation and a user

guide are available at http://pappulab.wustl.edu/CIDER/help/. The web server takes unformatted or FASTA formatted sequences as inputs. It uses an intelligent formatting algorithm to strip out non-alphabetic characters. The analysis performed by CIDER is synthesized in terms of a series of sequence-specific parameters and plots that quantify the information accessed from the sequence information that is input by the user. A sampling of the analysis that is provided by CIDER is shown in fig. 4.2. CIDER makes all the calculated sequence parameters available in downloadable text format. Multiple sequences can be analyzed and visualized simultaneously.

The CIDER webserver was written in the Python programming language using the Django web applications framework (https://www.djangoproject.com/). The user interface was built using the Bootstrap front-end framework (http://getbootstrap.com/). CIDER is deployed using an Apache webserver (http://httpd.apache.org/), running on OpenSuse Linux (https://www.opensuse.org/). No user information is stored and no information - other than usage statistics - are saved.

4.2.2 localCIDER

Unlike CIDER, localCIDER is a standalone software package that was developed to be a high performance, toolkit for the programmatic analysis of IDP sequences. It combines a wide array of sequence analysis routines with built-in plotting functions to create a single, all-encompassing framework for the analysis of IDP sequences. Installation information andaaa documentation are available via http://pappulab.github.io/localCIDER/.

The decision to create a standalone web server and a locally deployable software package was motivated by the fact that these two tools serve very different needs. A web server is

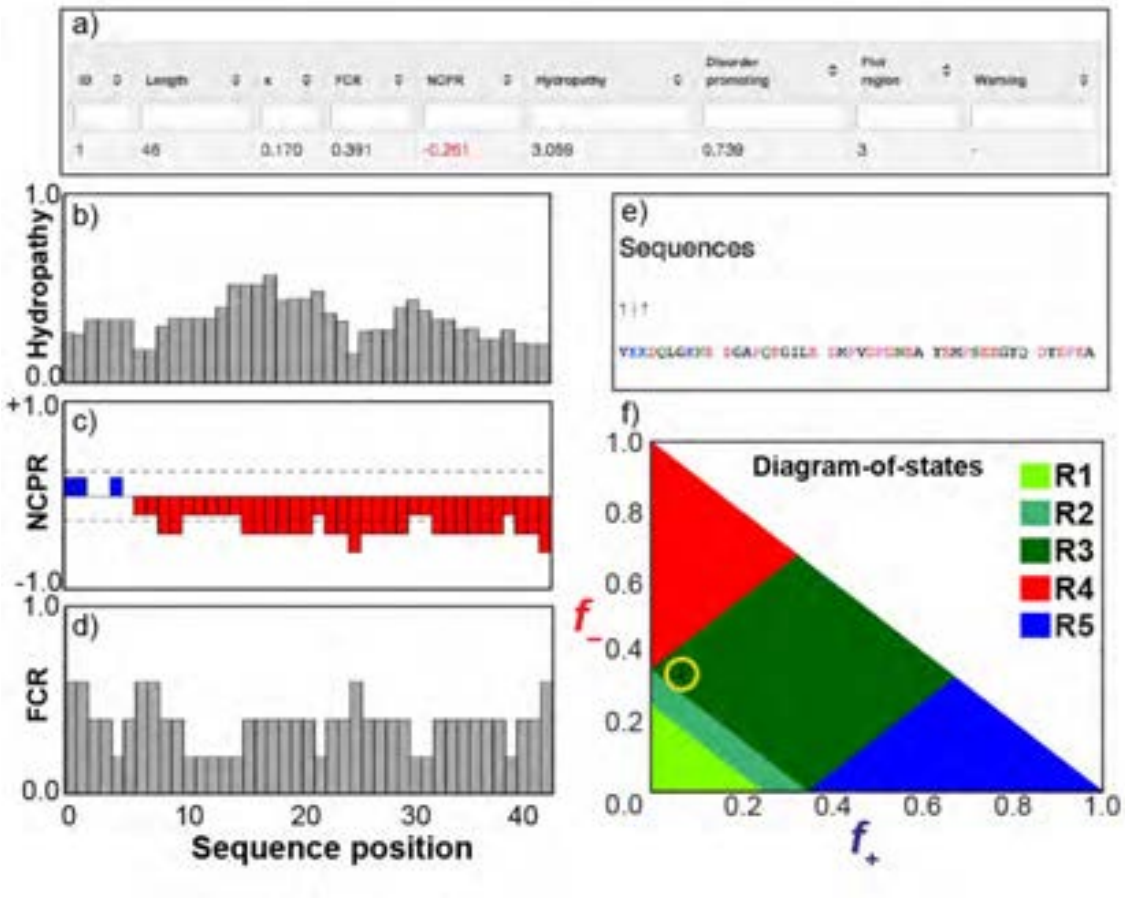

Figure 4.2: Overview of a subset of the output generated by CIDER. (a) shows an overview of the parameters that are calculated. When multiple sequences are analyzed each column is sortable. (b), (c), and (d) use a sliding window approach to show the linear hydrophobicity, NCPR, and FCR, respectively. (e) shows the physicochemically colored sequence. Here, black denotes hydrophobic residues, green denotes polar residues, and blue and red, respectively denote positive and negatively charged residues. (f) shows a sequence-annotated diagram-of-states.

ideal for quick, user-friendly access to summary statistics since this does not require any time or resource investments from the user. Web servers, however, introduce the complication of network latency, as well as providing a single point of failure when one seeks high-throughput

sequence analysis. localCIDER is an easily installable package that mitigates these issues and allows the deployment of powerful sequence analysis pipelines.

The localCIDER package implements a wide array of customizable analysis routines for the study of IDP sequences. There are two main classes of analyses in localCIDER: Sliding windows can be used to analyze local sequence features thereby generating position-specific descriptions of various physicochemical properties encoded by the IDP sequence. The sizes of sliding windows can be set to any value, which allows the analysis to be performed on any length-scale (see 4.20a). In addition, the user can quantify global descriptors that are computed as averages over the entire sequence. These include a range of parameters such as hydrophobicity, NCPR, FCR, κ, diagram-of-states classification, and average PPII propensities [115,161,541,582]. The local and global enrichment of particular classes of amino acids is readily visualized and quantified (see fig. 4.20b). The linear sequence complexity can be calculated using one of three possible complexity measures viz., Wootton-Federhen complexity, Linguistic complexity, and Lempel-Ziv-Welch complexity [321, 595, 651]. Many of these analysis routines allow the specification of user-defined adjustable parameters.

The parameter κ which quantifies the patterning of oppositely charged residues is calculated using a newly developed deterministic algorithm with O(1) complexity. If phosphosites are known a priori, these can be passed in as inputs and the distribution of κ values associated with various possible phosphorylation states can be calculated automatically, providing insight into how sub-stoichiometric phosphorylation would influence κ. Recently, we introduced a binary patterning parameter Ω that quantifies the linear mixing / segregation of proline and charged residues vis--vis all other residues. This is of particular relevance for high IDRs with high proline and low charge contents [366]. In addition to calculating Ω, can generalize the calculation of patterning parameters to any arbitrary binary sequence

patterning parameter. In this approach, one collects one set of residues into one group and all others into the second group. This allows one to investigate the mixing / segregation of any pair of residue types that are grouped into two categories. Examples include hydrophobic patterning, whereby all residues are grouped into hydrophobic or non-hydrophobic sets, disorder vs. order promoting residues, or neutral polar residues vs. all other residues. Similarly, analysis of ternary patterning, where residues are assigned to one of three groups, is also possible. Finally, input sequences can be converted into a reduced alphabet using either a set of pre-defined reduced alphabets or by passing in a user defined reduced alphabet mapping [407]. A reduced alphabet representation may be convenient for creating more coarse-grained sequence representations for further analysis, sequence clustering, and sequence comparison.

In addition to the various numerical analysis routines described above, all the linear analysis routines can directly generate pre-formatted PDF or PNG figures. In addition, diagram-of-state annotations and charge-hydropathy plots can be generated with an arbitrary number of different sequences on the same plot. The utility of an analysis package such as localCIDER comes from the ability to combine local and global sequence analysis with additional classification tools and statistical methods to enable rapid, customized high-throughput analysis pipelines, as demonstrated in chapter 11.

localCIDER runs on OSX, Linux and Windows, and requires minimal resource overhead. Plotting is carried out by matplotlib (http://matplotlib.org/) and numerical analysis by numpy (http://www.numpy.org/). localCIDER and its associated documentation are hosted freely on GitHub (https://github.com/), which is also used for version control and feature requests. For more information see http://pappulab.github.io/localCIDER/. A list of the full range of sequence analysis functions can be found at http://pappulab.github.io/localCIDER/

4.2.3 geeneus

Geeneus abstracts all network and data-parsing activity from the user. It works with a wide range of protein accession values (e.g. UnitProtKB, NCBI GI, RefSeq, etc.), automatically performing appropriate conversions and lookups when required. Session data are cached locally, meaning that for any given protein, a single network call is required regardless of the number function calls made. Data are returned in Python native formats for easy inclusion in larger-scale analysis pipelines. Geeneus and all associated dependencies can be installed from the Python package index using the '*pip*' tool. Geeneus was designed to give the illusion of having local, direct access to the full set of proteomic data housed in the NCBI, UnitProt, as well as calls to the Pfam database when necessary [176].

Geeneus is written in Python, with documentation available at http://alexholehouse.github.io/Geeneus/. The source code is freely available at https://github.com/alexholehouse/Geeneus. Geeneus has been in production for over two years, and is used in various applications, including the backend to the ProteomeScout webserver [370].

Geeneus boasts a number of features beyond the abstraction of all technical challenges from the user. A novel isoform reconstruction algorithm facilitates access to full isoform sequences from NCBI records. NCBI records store a single canonical sequence, and define isoforms in terms of changes relative to that canonical sequence. While formally complete, this definition makes obtaining the full-length sequence associated with a given isoform non-trivial. Our isoform reconstruction algorithm rebuilds the set of isoform sequences from this information. To ensure the assumptions made regarding the isoform annotation quality were founded, the algorithm was tested on 100,000 isoform-containing sequences. Without exception, all

isoforms annotations were found to be compliant in the standard required for the algorithm to function.

Geeneus operates via a data-memoization approach. When information regarding a protein of interest is requested, the complete dataset for that protein is fetched and stored in a local datastore. As a result, subsequent requests fetch data from a local in-memory cache, rather than across a network, although this cache can be refreshed on demand. This local store can also be exported to disk and then imported at a later date, meaning a savvy user requires only a single network operation per protein for any amount of analysis. Further, we have implemented batch fetching, which allows up to 100 accessions to be fetched simultaneously. Batch fetching uses a recursive divide-and-conquer retry procedure in the event that one or more of the accessions in a batch are invalid, meaning in all cases Geeneus runs the minimal number of network calls necessary to fetch all the data. NCBI request limits are hardcoded into the networking components to ensure compatibility with NCBI requirements.

4.2.4 PIUpred

To help provide a general framework for analysing and understanding disordered sequences we re-implemented the IUPred algorithm to perform high-throughput sequence analysis (correcting several minor bugs) [148, 149]. IUPred predicts disorder based on the local amino acid composition, and is parameterized from the PDB. In comparing meta-predictions (disorder predictions based on many different predictors) with equivalent predictions from IUPred we found extremely good agreement with high-confidence regions. We were asked not to re-distribute this new implementation, but the re-write is entirely in Python, providing an easy-to-use package that simply integrates into existing workflows. This allows us to *de novo*

predictor disorder in a high throughput manner (thousands of sequences per second). We will not discus PIUpred further here, but we introduce it only to make it clear that we have an in-house and highly efficient predictor, which has become an integral part of our sequence analysis workflow.

4.2.5 ProteomeScout& ProteomScout API

ProteomeScout is a database of protein-centric information, with a specific focus on post-translational modifications [370]. As well as contributing towards the tool itself, we developed toolkit (ProteomeScoutAPI,
https://app.assembla.com/spaces/proteomescout/wiki/ProteomeScoutAPI) to access and easily manipulate the large datasets associated with this data set [237]. Although post-translational modifications are not included in this chapter, we did find that highly charged sequence are more likely to experience a significant increase in κ upon phosphorylation.

4.3 Results

4.3.1 Obtaining Data for Proteome Wide Analysis

The following organisms were included in our full proteome analysis: *H. sapiens*, *R. norvegicus*, *M. musculus*, *G. gallus*, *A. thaliana*, *D. rerio*, *D. melanogaster*, *C. elegans*, *P. falciparum*, *D. discoideum*, *N. crassa*, *S. cerevisiae*, *S. pombe*, *C. albicans*, *B. subtilis*, and *E. coli*. All proteomic data were obtained from the UniProt reference proteomes, downloaded from the EBI FTP server (http://www.ebi.ac.uk/reference_proteomes). A list of these proteomes is

provided at the end of this subsection [600]. DisProt sequences were taken from the DisProt download (DisProt Release 7.03), which after redundancy filtering includes 744 disordered fragments of over 30 residues [455].

Disorder data for each proteome was taken from the MobiDB 2.0 consensus prediction data [461]. MobiDB combines disorder predictions from ten disorder predictors. A consensus prediction is generated as a majority vote based on those ten predictors, with a classification of '*disordered*' or '*structured*' assigned to each residue in each protein from the proteome. The result of this consensus disorder prediction was then post-processed to remove short islands ($\leq$ 3 residue) of disorder or order to create a less fragmented set of regions. Specifically, if an identified region - either a disordered region or a structured region - was found to be less than four residues long it was converted into the type of its surrounding regions. We compared results with and without this post-processing and found no difference in terms of the parameters reported in this study, though clearly this post-processing influences the number and size of IDRs identified, an aspect not examined in this work. The presence of short islands of order within disordered regions is primarily an artefact of combining multiple semi-overlapping predictors, as illustrated by fig. 4.3. The threshold of three or fewer residues was selected as a value of half the thermal blob length-scale (6-7 residues), i.e., substantially shorter than a length scale over which persistent structure would be expected [126].

The use of a consensus score, rather than relying on a single disorder predictor, helps to avoid any intrinsic biases in various predictors. It creates a more stringent threshold for defining a region as disordered, but ensures that, to the best of our ability, regions predicted as 'disordered' are utilizing approaches from multiple predictors to avoid false positives. In retrospective analysis we repeated much of the work done here using a single disorder

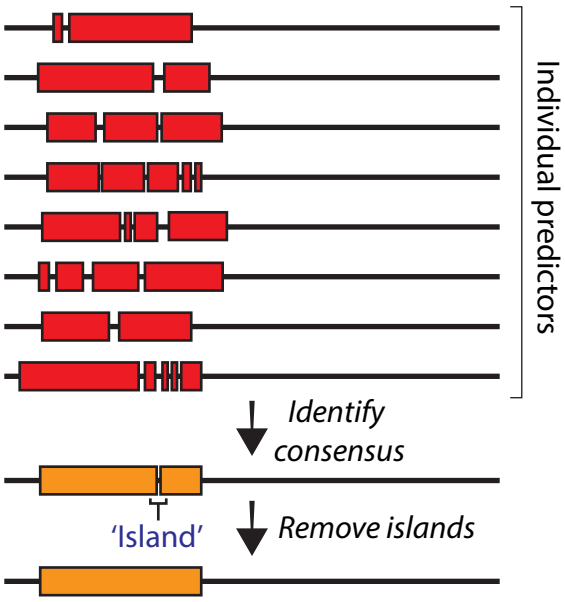

Figure 4.3: Schematic showing the creation of a consensus disordered region from multiple predictors, followed by the removal of a short 'order' island. Disorder predictors typically generate profiles with multiple short interruptions.

predictor (IUPred) and found highly analogous results (data not shown) suggesting that IUPred provides a robust stand-alone prediction [148].

MobiDB provides a consensus prediction based on a set of ten disorder predictors [461]. In addition to these ten predictors, MobiDB also allows for the inclusion of structural information from the Protein Data Bank (PDB) to further annotate structural preferences within a region. For our analysis, we used the MobiDB consensus data from disorder predictors alone, rather than also including additional information from the PDB. This decision was made based on two considerations. Firstly, many IDPs are known to undergo coupled folding and binding. The PDB contains a large number of structures representing protein regions that have been shown to fold in the context of a partner, and while relevant for function, this

does not appear to be relevant for knowing the region's intrinsic structural propensity as an autonomous unit. As a result, MobiDB's approach of using structural data to categorically rule out a region as disordered is highly appealing, but may unintentionally yield false negatives in some circumstances. As a specific example, the protein PUMA (p53 up-regulated modulator of apoptosis) is predicted to be disordered, and yet the mouse variant (UniprotID Q99ML1) contains a region (residues 130-155) that has been structurally characterized by NMR (PDB ID 2ROC) [132]. As a result of this apparently alpha-helical region, MobiDB defines this region as structured (fig. 4.4A). However, this region adopts a stable helix only upon binding to its partner Mcl1 (fig. 4.4), and has been experimentally shown to be disordered in the unbound state, although recent studies show a roughly 40% likelihood that PUMA adopts helical conformations in its unbound form [221, 494].

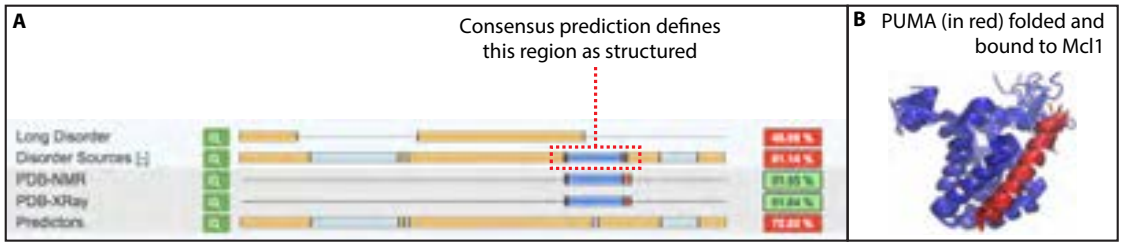

Figure 4.4: Example of structural data incorrectly informing on a folded region. Panel A shows a screenshot from the MobiDB website, and highlights the fact that the full consensus prediction approach used assigns the region in blue to be folded. Panel B shows the NMR structure of PUMA bound to Mcl1 (PDB ID 2ROC).

Given that the analysis carried out in this study compares multiple proteomes, we felt that it was important to use a uniform approach across all the primary (sequence) data. If structural data were included, we would intrinsically bias sequences from organisms that have been studied in greater structural detail, towards being more likely to identify structured regions. This could have the unintended consequence of allowing a region to be classed as disordered in one organism but structured in another if structural data had been obtained in one species but not in the other.

Having obtained the set of disordered regions associated with a proteome, each disordered region greater than thirty residues and with a proline content of less than 15% was used for further analysis. A threshold of thirty residues was chosen to match the general consensus of 'long' disorder [148]. The threshold of 15% for proline content is in keeping with the original definition of the diagram-of-states, and the fact that a growing body of evidence suggests that enrichment in proline drives ensembles to be more expanded than one might navely expect based on FCR alone [126, 359, 364, 405]. The influence of proline residues is explored systematically in other work, and the patterning of charged and proline residues is quantified by the parameter Ω (see chapter 6).

Proteome-wide statistics for various quantities are shown in table 4.3.1. The "percentage 'long' disorder" represents the percentage of the proteome from each organism which is encompassed by a single disordered region stretching thirty residues or longer. Other estimates in the literature do not use this 30 residue threshold (and use a less stringent disorder classification), and as such find a much higher percentage of disorder in the proteomes from these organisms.

Species	**Number of PBD structures**
Homo sapiens	37183
Rattus norvegicus	2612
Mus musculus	6834
Gallus gallus	1635
Arabidopsis thaliana	952
Danio rerio	202
Drosophila melanogaster	905
Caenorhabditis elegans	306
Plasmodium falciparum	649
Dictyostelium discoideum	153
Neurospora crassa	69
Saccharomyces cerevisiae	4485
Schizosaccharomyces pombe	334
Candida albicans	123
Escherichia coli	13538
Bacillus subtilis	1510
Thermotoga maritima	655

Table 4.1: Summary of the number of structures in the PDB by species

Reference proteome	Organism	Number proteins	# IDRs	% dis.
UP000005640_9606	Homo sapiens	20 882	23 437	18.60%
UP000002494_10116	Rattus norvegicus	21 866	21 529	17.40%
UP000000589_10090	Mus musculus	22 129	22 448	17.30%
UP000000539_9031	Gallus gallus	15 749	16 949	16.60%
UP000006548_3702	Arabidopsis thaliana	27 221	17 192	11.50%
UP000000437_7955	Danio rerio	25 642	25 125	16.50%
UP000000803_7227	Drosophila melanogaster	13 674	15 489	19.80%
UP000001940_6239	Caenorhabditis elegans	20 274	12 716	13.10%
UP000001450_36329	Plasmodium falciparum	5 162	7 274	14.60%
UP000002195_44689	Dictyostelium discoideum	12 732	13 703	20.30%
UP000001805_367110	Neurospora crassa	9 756	11 927	23.90%
UP000002311_559292	Saccharomyces cerevisiae	6 720	5 381	14.60%
UP000002485_284812	Schizosaccharomyces pombe	5 104	3 407	11.70%
UP000000559_237561	Candida albicans	8 354	6 450	16.60%
UP000000625_83333	Escherichia coli	4 305	274	1.15%
UP000001570_224308	Bacillus subtilis	4 197	382	1.75%
UP000008183_243274	Thermotoga maritima	1 851	47	0.42%

Table 4.2: Summary proteomic statistics relevant for this study. Note that the **# IDRs** refers to the number of disordered regions that are 30 or more residues in length, and the **% dis.** reflects what percentage of the proteome falls into these 'long' IDRs.

These data represent a total of 243,644 proteins, 203,683 disordered regions, and an average "percentage long proteome disorder" of 16.6% in eukaryotes and 1.45% in non-hyperthermophilic prokaryotes (*E. coli* and *B. subtilis*). The significant depletion of long disordered regions

in the hyperthermophile *T. maritima* is in line with previous work [80]. With so few disordered regions in *T. maritima*, it was excluded from further analysis in this study to avoid the introduction of misleading biases.

4.3.2 Proteome Wide Analysis of Disorder

We first examined how disordered regions are distributed across the diagram-of-states (fig. 4.5A). For the human, rat, mouse, and chicken proteomes the distribution across R1-R5 was highly similar, and generally matched the DisProt distribution. For other organisms (notably *D. melanogaster*, *P. falciparum*, and a number of fungi) large deviations from the distribution seen in humans were observed. In all cases, relatively few polyelectrolytes (R4/R5) were identified, and those found were almost exclusively negatively charged. We also found that the fraction of charged residues (FCR) varied between different organisms (fig. 4.5B), as do the distributions of κ values (fig 4.5C). Taken together, these results show that the distribution of charge density and patterning vary across organisms, although similar global trends are also observed. An additional takeaway from this analysis is that by these measures, DisProt encompasses a good representation of IDPs for describing the sequences in the human proteome. One could have imagined that DisProt might have been enriched in charged IDPs, but this analysis firmly shows that DisProt provides a representative snapshot of the human IDPs, at least in terms of amino acid composition.

Having established that the median FCR for disordered regions varies across different organisms, we asked if the distribution of κ values observed for naturally occurring sequences varied with FCR. To answer this question, we focused on polyampholytic sequences (absolute net charge per residue $|\text{NCPR}| < 0.25$, $\text{FCR} > 0$). Based on anecdotal evidence, κ appears

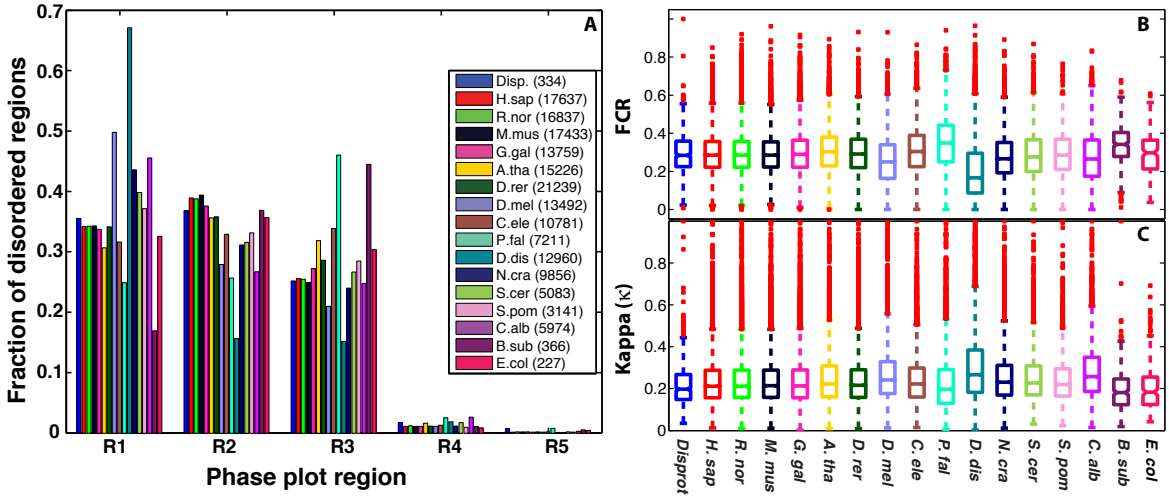

Figure 4.5: Sequence properties of disordered regions across sixteen proteomes and DisProt. Legend numbers indicate the total number of disordered regions identified. Panel A shows fractional populations of the diagram-of-states regions for all IDPs from sixteen different model organisms and the DisProt database. While broadly similar trends are observed, there are substantial differences between different organisms. Panel B is a box-plot showing the distribution of FCR values for all IDPs taken from the same set of organisms and DisProt. The central box defines the first quartile, median, and third quartile from the data. Similarly, Panel C is a box plot showing the distribution of κ values taken from the same set of organisms and DisProt.

has the most significant influence on the conformational behavior of sequences which display the dual traits of an intermediate-to-high FCR and a near neutral overall charge - i.e., strong polyampholytes. For comparison, we generated a random prior model by taking each disordered region and performing a fully randomizing shuffle of the sequence. To facilitate the generation of such a background, an efficient method for performing sequence shuffling is implemented in localCIDER. This process generates a composition and size-matched dataset

with identical FCR and NCPR distributions, but where the κ of each sequence has been altered. By constructing a random prior we can examine how κ-varies as a function of FCR in the absence of any selective pressure for sequence patterning. This is an oversimplification given the fact that many other residues show local sequence compositional preference, but is a simple and consistent approach to generate a conceptually important random prior.

Fig. 4.6A shows a comparison of median FCR vs. median κ across the different organisms. The statistically expected behavior obtained from the composition matched random prior is that κ should be inversely correlated with FCR, as shown by the red dashed line. In naturally occurring sequences we found a strong inverse correlation between κ and FCR with a steeper gradient than would be expected from randomly shuffled sequences; the gradient for the random prior (red dashed line) is 0.24, while the gradient for the naturally occurring sequences (black solid line) is -0.54.

In further analysis, we examined the 2D probability distribution of FCR and κ for all species relative to the same random background (fig. 4.6B). We found a significant over-representation of intermediate-κ and intermediate-FCR disordered regions in naturally occurring sequences (red region). Throughout naturally occurring sequences we found an absence of high κ / high FCR sequences, in line with anecdotal experimental results where highly charged sequences with a high κ are often aggregation prone due to strong electrostatic attraction between oppositely charged patches.

4.3.3 FCR vs. κ - Further Analysis

Fig. 4.6B shows a two-dimensional density difference plot, generated by creating two 2D histograms (fig. 4.7) and subtracting the random distribution from the naturally occurring

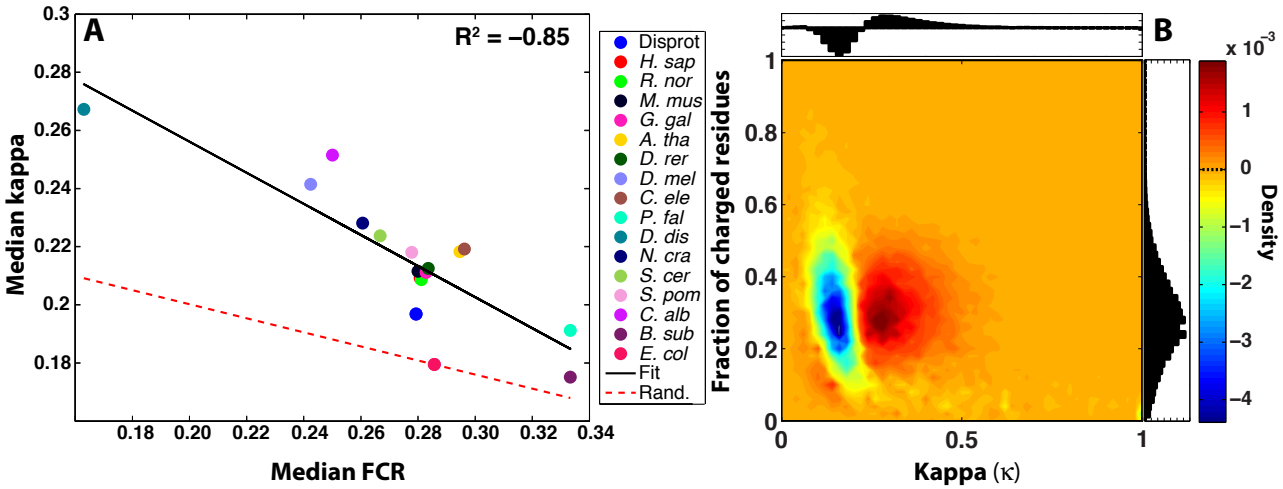

Figure 4.6: Panel A examines the relationship between median FCR and κ across the different organisms. Panel B is a 2D histogram difference map, and shows regions that are enriched (hotter colors) or depleted (cooler colors) in naturally occurring sequences with respect to a randomly scrambled composition matched background set of sequences. We found that naturally occurring sequences are enriched for sequences with higher κ values, suggesting the evolutionary selection for charge segregation.

distribution. As described previously, for this analysis (κ vs. FCR) we focused on polyampholytic sequences. The raw 2D distributions used to generate Fig. 4.5B are included below. In these 2D histograms a bin size of 0.02 is used for κ and for the FCR.

Fig. 4.8 shows the distribution of κ values, comparing all naturally occurring sequences with the random-prior sequences. As expected based on the 2D distributions shown in fig. 4.6 and 4.6, we find that naturally occurring sequences show a broader distribution of κ-values. Notably, substantially more sequences have a higher κ-value than would be expected from a random distribution. Again, this result is in line with naturally occurring sequences having more 'charge blocks' - local regions of high net charge density - than one would expect by random chance.

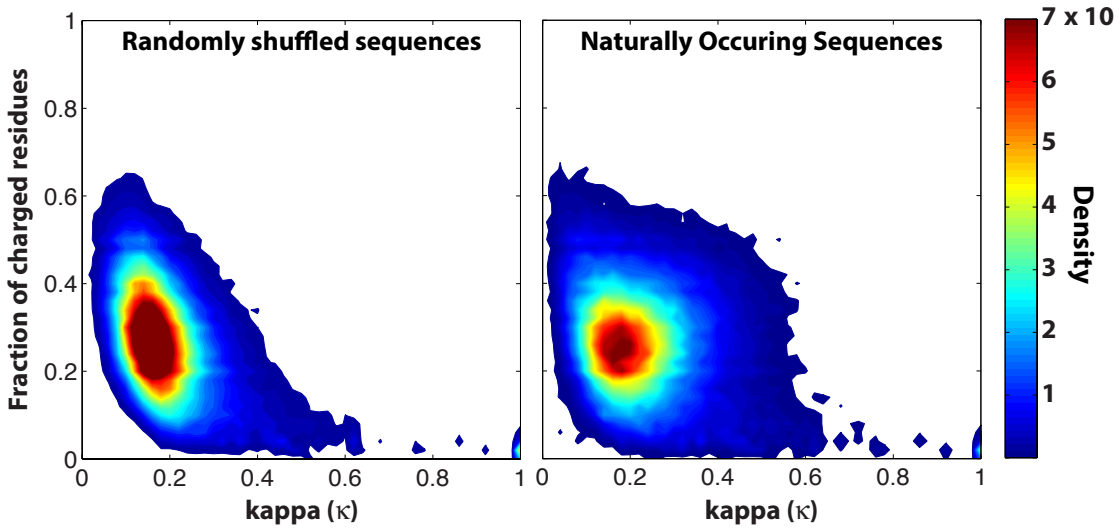

Figure 4.7: 2D histogram showing the density of sequences associated with a specific κ value and FCR for all polyampholytic disordered regions. fig. 4.7B is the differences between these two 2D histograms.

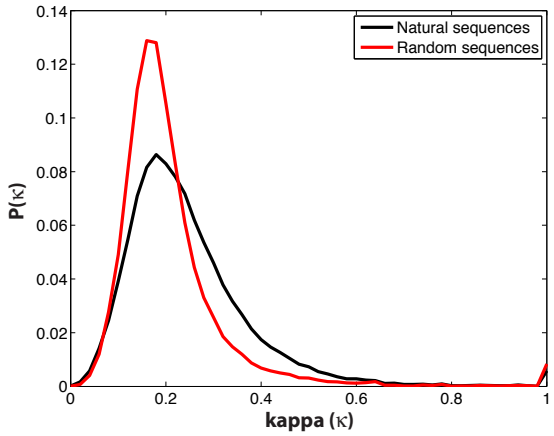

Figure 4.8: Histogram generated distributions of κ-values from naturally occurring sequences (black) and from the same sequences after unbiased sequence scrambling (red)

The data in fig. 4.7 can also be represented by determining the median FCR values for a specific κ-range and considering how the median FCR varies with κ. In this analysis, as

shown in fig. 4.9, the complete set of naturally occurring sequences are sub-divided into bins based on κ. For the sequences in each bin, the median FCR value is calculated, and the median FCR vs. κ-range is plotted. This analysis does not provide information regarding the number of sequences associated with a given FCR, but allows different organisms to be compared with one another in terms of how FCR and κ co-vary. Only bins with 50 or more sequences are plotted. This analysis reproduces the trends observed in fig. 4.6A the median κ and median FCR are inversely correlated with one another. Again, this implies that there are few sequences where κ and FCR are simultaneously high. We found that sequences with a low κ-value are strongly biased towards a high charge fraction, whereas sequences with a high κ-value are generally depleted in charged residues. Beyond these observations, extracting meaningful proteome-wide conclusions from these data is difficult. While charge distribution described by κ is an important component in determining an IDP's ensemble, there are many other contributing factors that vary on a case-by-case basis. As a result, these data provide a general, big picture summary of the expected trends, but over-interpretation should be avoided.

An important yet nuanced observation from the data presented thus far is that for each organism there exists a range of FCR values where a wide variety of κ - values are observed; for many organisms this FCR range lies in the interval of $\sim$0.2 to $\sim$0.4 (fig. 4.9). This overlaps with the R2 region on the diagram-of-states (fig. 4.2), one of the regions (along with R3) where κ has the greatest influence on conformational properties. Finally, R2 is generally the region on the diagram-of-states with the greatest number of disordered regions (fig. 4.5a). Taken together, these results suggest that a significant fraction of naturally occurring IDPs taken from a wide range of different organisms display sequence properties where charge patterning would be expected to play a major role in determining their conformational ensemble. This is a necessary but not sufficient result to assert that charge patterning is an

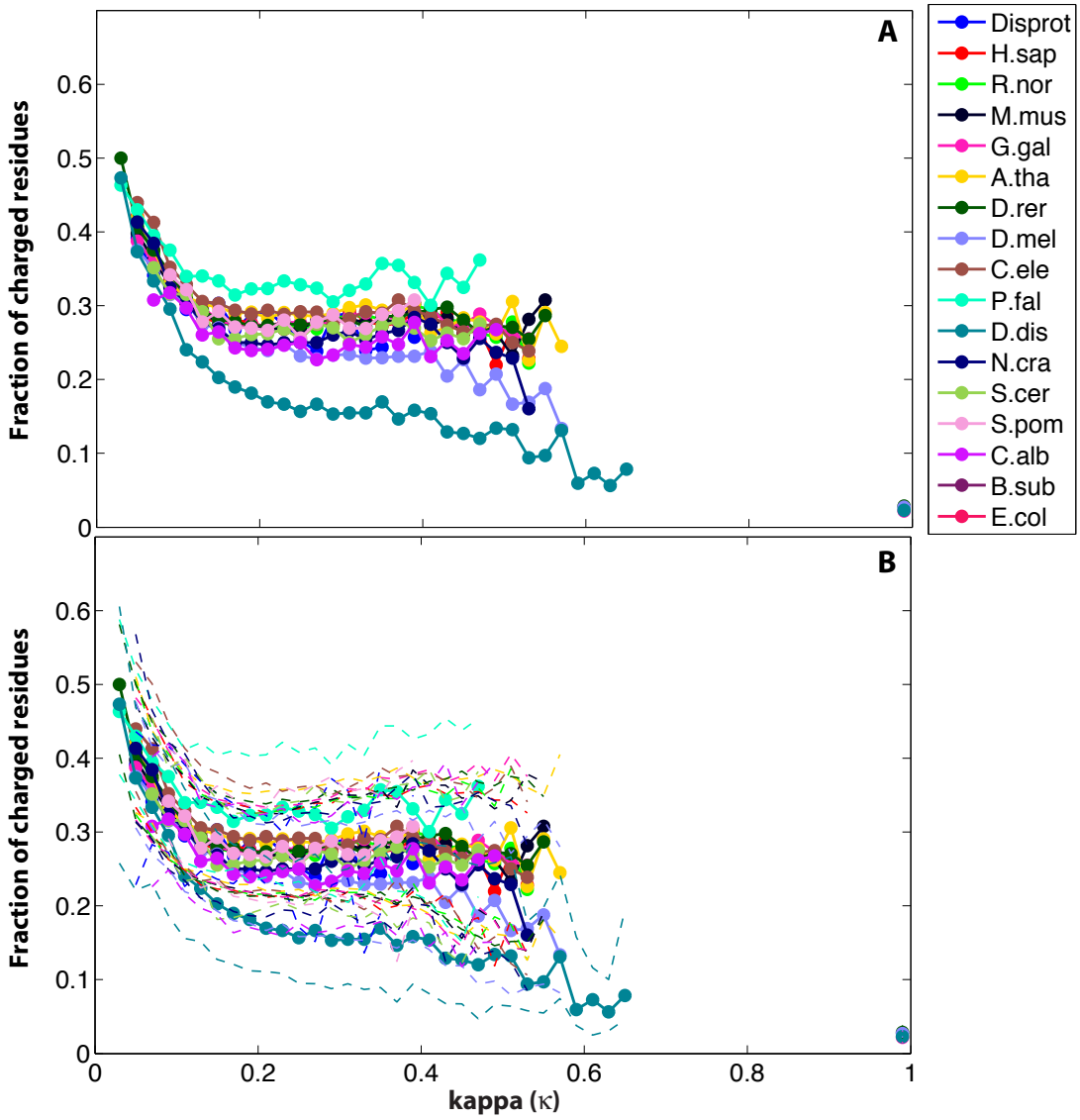

Figure 4.9: Binned FCR representation of the relationship between κ and FCR. Panel A shows the data without the interquartile range bounds, while Panel B shows the same data with interquartile range bounds.

important feature for proteins from many different organisms, but sets the stage for a deeper investigation into specific examples through higher resolution analysis.

Given the preceding discussion combined with the inverse correlation between FCR and κ, we offer a plausible interpretation of our results. For sequences with a high FCR there appears to be a strong selective pressure towards well mixed sequences (low κ). At intermediate FCR ($0.2 \leq \text{FCR} \leq 0.4$) sequences experience a range of selective pressures both for lower than expected and higher than expected κ values giving rise to a wider distribution than would be expected in the absence of any selective pressure. Finally, at low FCR ($0 \leq \text{FCR} \leq 0.2$) charge patterning becomes less influential, and as a result these sequences experience weaker selective pressure. For a true assessment of the conservation associated with sequence patterning, an analysis should consider paralogous IDRs across many different species. While not examined here, clearly localCIDER is well placed to aid in this kind of sequence analysis. It is also worth emphasizing that many other factors (conserved recognition motifs, amino acid composition, residual secondary structure) will all play a role in determining the evolutionary landscape, although charge patterning as measured by κ are one pair of relevant sequence features.

4.3.4 FCR vs. NCPR

In the previous section we examined proteome-wide distributions of κ and FCR. Analogously, we can examine how NCPR varies with FCR. Fig. 4.10 shows the 2D histogram - using the same approach as in fig. 4.4 of FCR vs. NCPR. To maintain bin number parity, the NCPR bin size is 0.04 (ranging from -1.0 to 1.0) while the FCR bin size remains at 0.02 (ranging from 0.0 to 1.0). For these analyses we did not filter out any polyampholyte sequences, but instead used all available sequences.

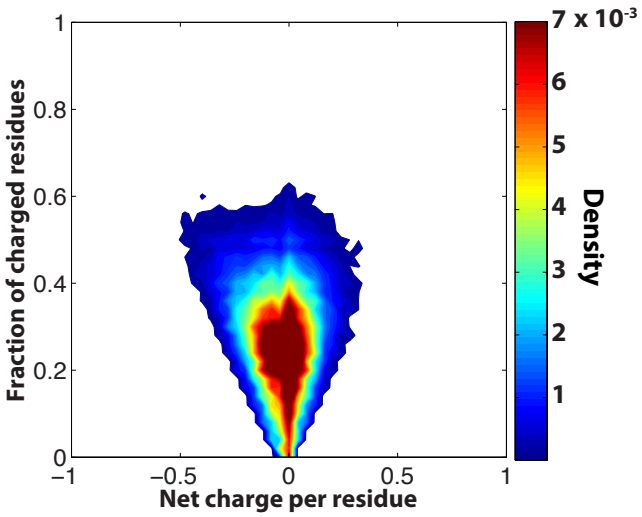

Figure 4.10: 2D histogram of FCR vs. NCPR. We find the majority of disordered regions are polyampholytes, with an FCR of between ∼0.18 and 0.35.

This analysis shows that, generally speaking, there is a depletion of polyelectrolytes across the discorded regions in naturally occurring proteomes, as observed in fig. 4.5A. Fig. 4.11 shows how NCPR is distributed across the different organisms.

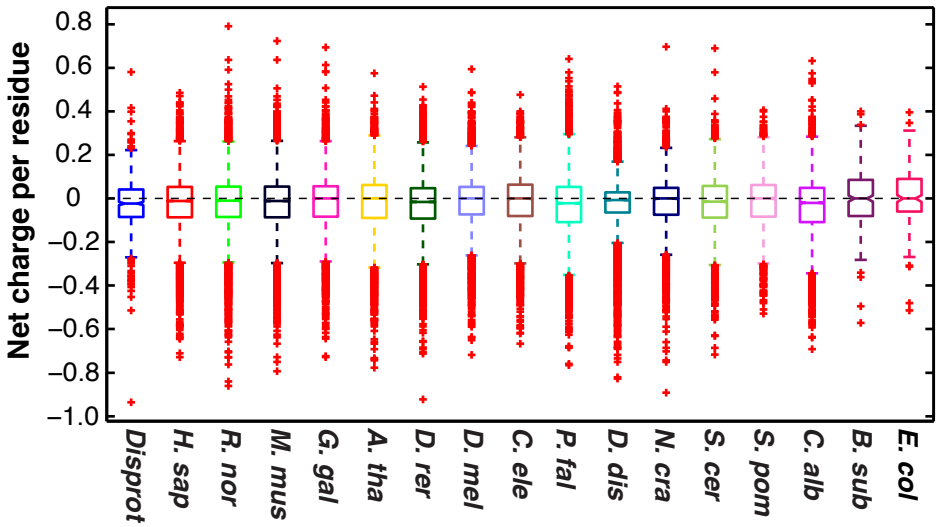

Figure 4.11: Net charge per residue (NCPR) distribution across organisms. Similar trends are observed over the wide variety of organisms examined.

We can further examine this distribution using the sequence binning approach employed in the FCR vs. κ analysis in fig. 4.9. Fig. 4.12 shows that the same trends with respect to charge are observed across all organisms as the FCR of regions increases, the NCPR tends towards being increasingly negative (acidic).

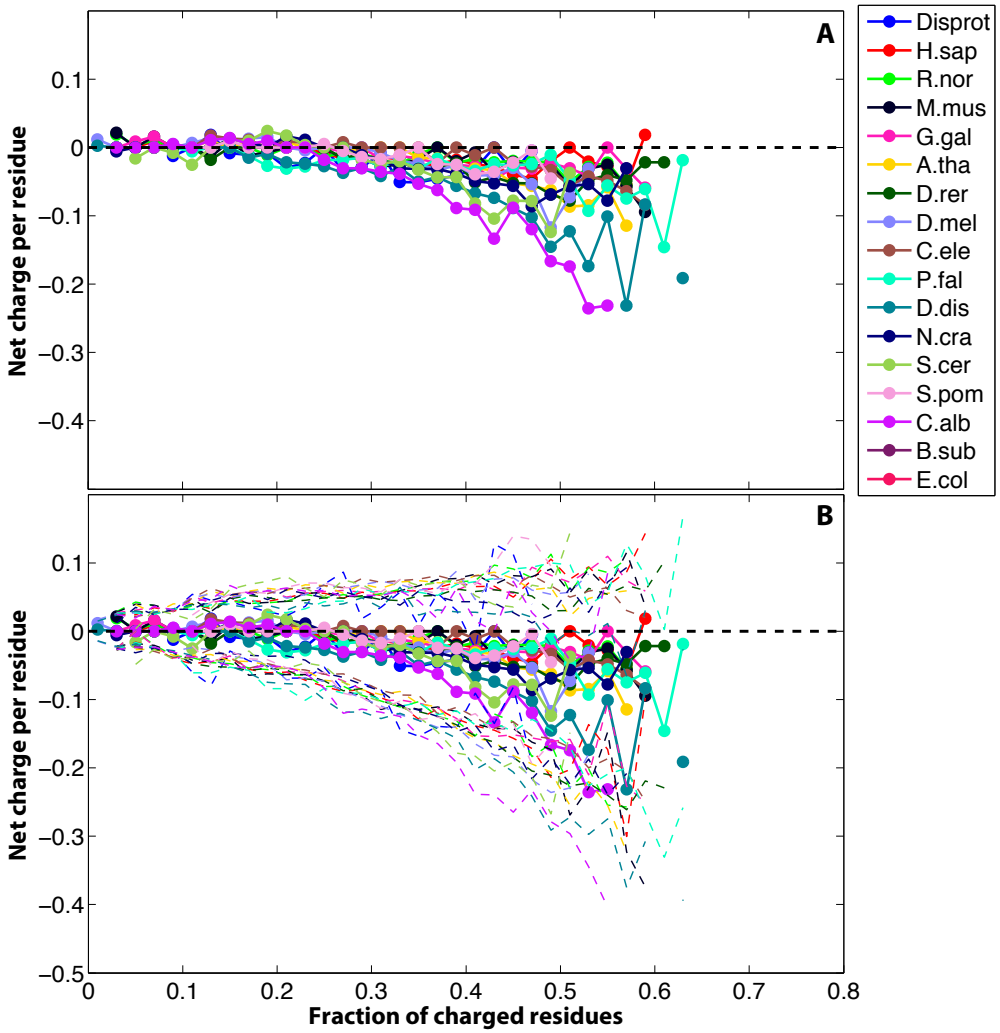

Figure 4.12: The consistent trends observed suggest that more highly charged disordered regions will have a modest net negative charge. Some organisms (e.g., *C. albicans* or *P. falciparum*) contain disordered regions that are more strongly acidic at a high FCR. Panel A shows the median NCPR when sequences are binned according to their FCR values. Panel B shows the same information as panel A, and includes the interquartile range as dashed lines. The thick black dashed line represents the neutral NCPR value.

An interesting observation that emerges from these data is that sequences with a high fraction of charged residues are most likely to carry a net negative charge. Moreover, there is a near total absence of highly charged sequence with a net positive charge observed across all organisms. An important caveat to consider in all this analysis is that we make no attempt to de-convolve disordered regions into sub-domains, such that all properties examined are the average over each contiguous disordered region. Given the distinct conformational preferences associated with IDPs of different sequence composition, we have no reason to assume that disordered regions could not be divided into sub-domains, where long disordered regions (e.g. > 200 residues) may contain functionally and conformationally discrete subdomains. The identification of such domains represents a future goal that will be achievable using localCIDER, but is beyond the scope of this work.

4.3.5 A Discussion on κ

For completeness, we provide a detailed discussion on the statistical and practical properties of the patterning parameter κ (kappa). This section is divided into several subsections. In section 9.1 we revisit how κ is defined, providing an intuitive overview combined with a mathematical definition. In section 9.2 we provide a method to formally calculate the number of charge permutants, and describe the probability mass function (PMF) associated with κ for a given sequence composition. Section 9.3 describes how the expected κ value varies across the diagram-of-states with complete enumeration of expected values for all possible compositions over a range of different sequence lengths. Finally, in section 9.4 we offer some general rules of thumb when thinking about how a sequence' κ value may influence conformation. Section 9.4 also offers some notes of caution regarding how one should or should not treat the parameter.

Defining κ

The ideas presented in this subsection were first described in previous work [126]. We include them here for completeness. The parameter κ is a measure of the mixing of oppositely charged residues along the primary sequence of a protein, where this mixing is effectively quantifying how similar the local charge distribution is when compare to the global charge distribution. Specifically, the local charge distribution is assessed based on five and six residue sub-fragments (blobs). For a sequence where charged residues are globally well mixed with respect to one another, local sequence properties and global properties will mirror one another. For a sequence where charged residues are highly segregated, local properties will be consistently divergent from the global properties. κ is a parameter that formally describes this similarity/difference, and is normalized against a maximally segregated sequence to ensure $0 < \kappa \leq 1$. A graphical summary of how κ maps to protein sequences is shown in Fig. 4.13.

To compare local vs. global properties, it is necessary to define a comparison metric. Such a metric should be normalized by sequence length to allow the comparison of regions of different lengths (e.g., a local six residue blob compared with the full n residue sequence). For κ, the metric used is charge asymmetry (κ), which is defined as follows

$$\sigma = \frac{(f_+ - f_-)^2}{(f_+ - f_-)} \tag{4.1}$$

Here, f_+ and f_- represent the fraction of positively and negatively charged residues. To carry out a complete comparison of the global sequence properties with the local sequence properties, we perform a comparison of all possible blobs with the full sequence, normalized

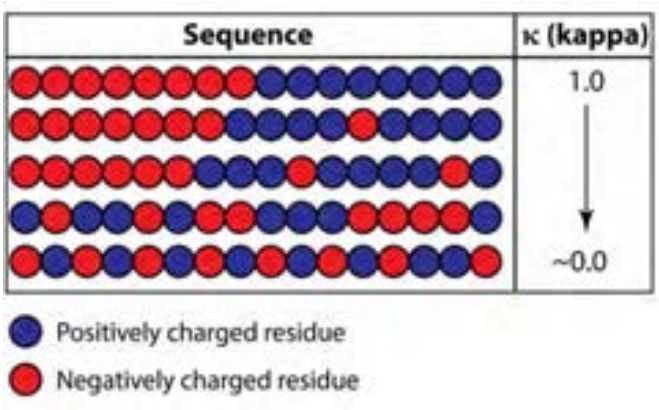

Figure 4.13: Graphical description of how κ-varies with sequence patterning. A high κ-value is associated with a highly segregated sequence, while a low κ-value is associated with a well-mixed sequence. It is worth noting that were are using highly charged polyampholytes here because they graphically illustrate the relationship between κ and patterning well, but as a relevant parameter, κ also applies to much less charged naturally occurring sequences.

by the number of blobs. Specifically, each blob is g residues long, meaning a sequence of n residues is subdivided into $(n - g + 1 = N_{\text{blobs}})$ blobs. For a complete comparison of global vs. local properties we introduce a new parameter, (δ) which defines a permutant-specific comparison between global and local charge asymmetry, and is defined by:

$$\delta = \frac{\sum_{i=1}^{N_{\text{blobs}}} (\sigma_i - \sigma)^2}{N_{\text{blobs}}} \tag{4.2}$$

Here, σ defines the charge asymmetry for the full sequence while σ_i defines the charge asymmetry associated with the i-th blob. A graphical schematic of how the summation terms in are calculated is shown in fig. 4.14 (where $g = 6$);

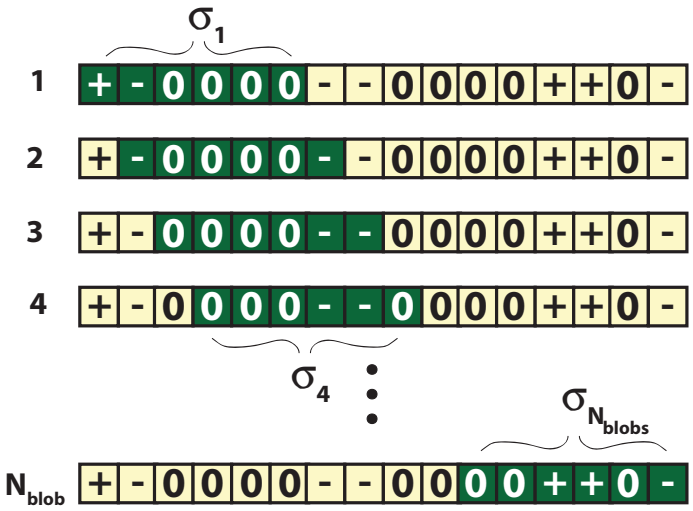

Figure 4.14: Graphical schematic showing how the summation term in the δ calculation represents a sliding window for determining the σ for each overlapping blob of g residues (where in this case $g = 6$).

Having calculated δ for the sequence of interest, we introduce a normalization factor to ensure that we have a parameter (κ) that ranges from 0 to 1. This normalization factor ($\delta_{\max}$) represents the δ associated with the maximally segregated sequence, such that we define κ as shown below

$$\kappa = \frac{\delta}{\delta_{\max}} \tag{4.3}$$

Finally, for the full definition of κ we need to define the blob size (g) i.e., what is the length-scale that we consider 'local'. The value selected for g was chosen to reflect the number of residues that give rise to a chain length at which the interplay between chain-chain, chain-solvent, and solvent-solvent interactions are on the order of kT. This refers to the thermal blob, as discussed in more detail in section 13.2.5 [146]. For protein sequences

with low proline contents (i.e., less than 15%) this value is 5 to 6 residues. To account for this variability, we use an average of the κ-value derived from g = 5 and g = 6. As a result, the κ value reported in the original paper and by localCIDER and CIDER is defined by equation

$$\kappa = \frac{\frac{\delta^{g=5}}{\delta^{g=5}_{\max}} + \frac{\delta^{g=6}}{\delta^{g=6}_{\max}}}{2} \tag{4.4}$$

Where $\delta^{g=i}$ reflects the value calculated for a sequence with a blob size of i.

Number and distribution of κ-values

Having defined how κ is calculated, it is useful to provide a general sense of the range of κ values that are likely, given a sequence composition. The most intuitive approach to answer this question would be to define a probability mass function (PMF) associated with κ for a given sequence composition. This would provide a statistical description of the likelihood associated with a given κ-value, and help offer statistical context for the κ-value associated with a naturally occurring sequence i.e., is it far from or close to the statistically expected value. In the following subsection we examine how κ values are distributed, and how this distributions changes with FCR and NCPR. An important point to reiterate is that many different sequence permutants will have the same value of κ, a consequence of the fact that κ is a scalar parameter trying to capture sequence-encoded patterning.

One approach for generating the κ PMF would be to perform exhaustive enumeration and determine the complete mapping of every possible charge permutant to κ value followed by the creation of a histogram of those κ values. The number of possible charge permutations

of a sequence can be calculated by taking the sequence, converting it into a three-letter alphabet representation (negative, neutral, positive) and using the expression defined below

$$\begin{aligned} \text{Number of permutations} &= \binom{n_0 + n_+ + n_-}{n_0}\binom{n_+ + n_-}{n_+}\binom{n_-}{n_-} \\ &= A \times B \times C \end{aligned} \tag{4.5}$$

In this expression, n_0, n_+, and n represent the number of neutral, positive, and negative residues, respectively. The notation here shows the product of three "*a choose k*" terms (termed A, B, C for convenient discussion below). For an example of the conversion of an amino acid sequence into a three-letter alphabet representation, see the twenty residue example in fig. .

Figure 4.15: Example of converting a twenty residue peptide from to a three-letter alphabet. This peptide's sequence-properties are $n_0 = 9$, $n_+ = 5$, and $n = 6$, giving it an FCR of 0.55 and an NCPR of -0.05.

Equation 4.5 can be explained by considering the following framework. There are a total of $(n_0{+}n_+{+}n)$ positions in the sequence. n_0 of those positions can be filled by neutral residues in A ways. This leaves $(n_+{+}n_-)$ positions, which can be filled with positive residues in B ways. Finally, there is only one permutation of ways the negative residues can fill, hence $C = 1$. For a 50-residue sequence with 7 positive and 7 negative residues (FCR = 0.28, NCPR = 0.0) there are approximately 3.2×10^{15} different permutations. For a 100-residue sequence with

20 positive residues and 20 negative residues, there are approximately 1.9×10^{39} different permutations. Based on these numbers it should be clear that complete enumeration of unique sequences and calculation of the associated PMF is not a feasible strategy. However, given the multinomial nature of the number of permutations, the distribution of κ values can be fit to a log-normal distribution. To illustrate this, we took all the disordered regions from the human proteome and generated 500 random permutants per region (i.e., $500 \times 23437 = 11718500$ sequences). For each random permutant we calculated the κ value. Consequently, for each of 23,437 IDRs we have a distribution of κ values generated through random shuffling. For each region, the distribution of κ values was then fit to a log-normal probability distribution, and the goodness of that fit assessed based on the Euclidean distance between empirical histogram and the log-normal fit. A schematic of this process is illustrated in fig. 4.16.

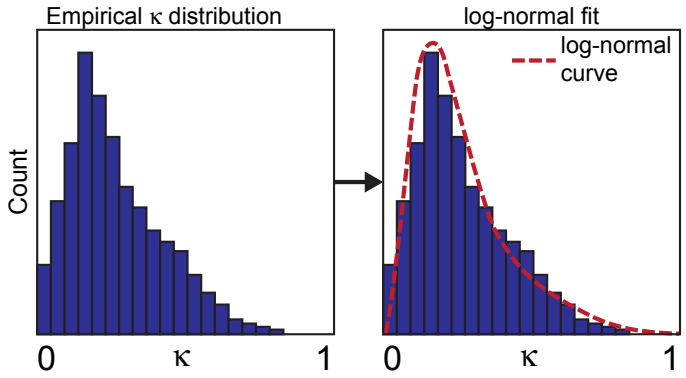

Figure 4.16: The panel on the left is an empirical histogram of κ values generated by random shuffling of a single IDR. The panel on the right shows a log-normal fit to that histogram (red dashed curve). The goodness of this fit is evaluated by determining the Euclidean distance between the empirical distribution and the log-normal distribution.

With the exception of sequences with a very low fraction of charged residues, we found that the log-normal distribution offers an extremely good fit to all possible regions. Fig. 4.17 shows the goodness of fit plotted in the diagram-of-states plot space (higher numbers indicate a greater deviation from the log-normal curve i.e., lower is better).

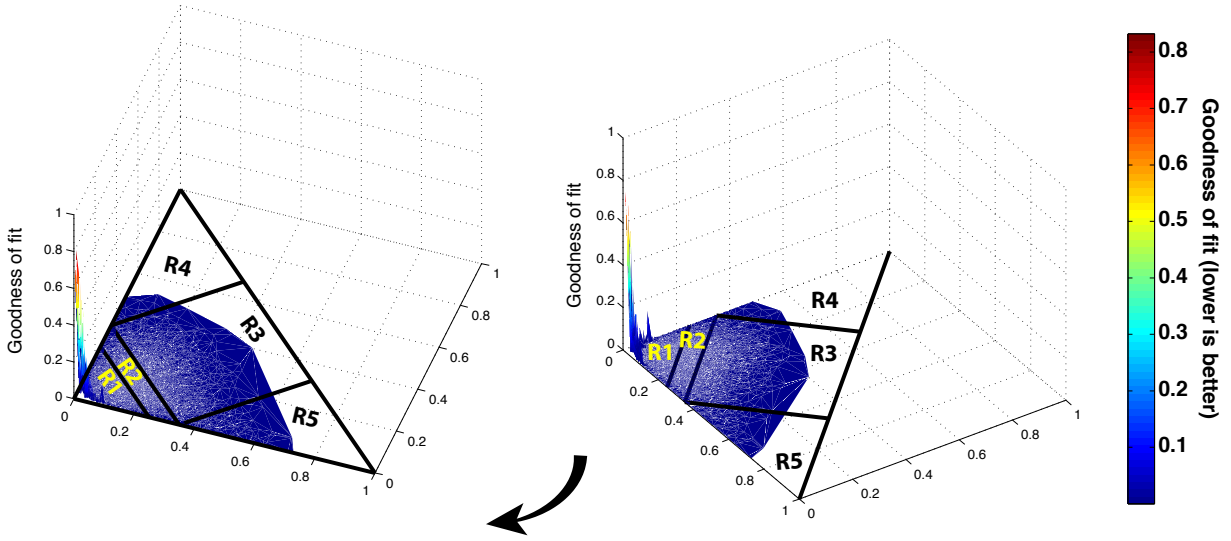

Figure 4.17: The goodness of fit of the distribution of possible κ values to a log-normal function is shown for all disordered regions in the human proteome shown as a 3D density plot superimposed on the diagram-of-states. The only regions where the fit does poorly is where FCR < 0.1 i.e., where κ stops being a useful parameter.

To better demonstrate the versatility of the log-normal fit, we randomly selected thirty examples of disordered regions, with six that that were of length 50, 75, 100, 200, and 300 residues. The empirical histogram vs. the log-normal fit is plotted in fig. 4.17. The goodness of fit across a wide range of lengths quantifies the robustness of the log-normal distribution as a reasonable approximation for the true distribution. Given the fact that the distribution of κ values can be approximated by a log-normal function it is now possible to determine

the true random likelihood of realizing the κ value of a naturally occurring sequence, i.e., we can ask "What is the probability that a sequence with a specific composition will have the observed κ value by random chance?". This analysis could help identify sequence where the κ value is far from the expected value, implying evolutionary pressure towards a specific κ value.

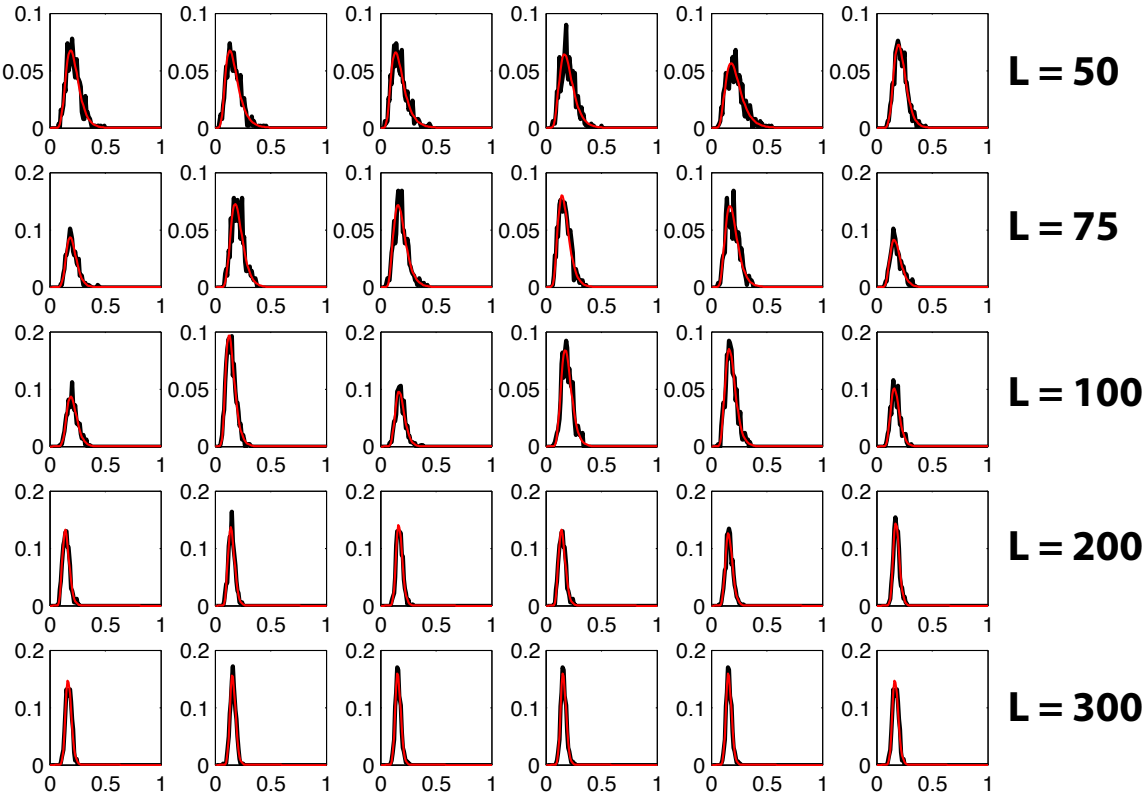

Figure 4.18: Randomly selected disordered regions and their empirical distribution of κ values (generated by determining the κ value of 500 random permutations of the sequence) compared with a fit log-normal distribution. In each subplot the abscissa (x-axis) is the κ value and the ordinate (y-axis) is the probability of that κ value. Each row contains six randomly selected sequences of length L=X (as defined in the far right hand side of the row). Red curves describe the log-normal fit, while black curves are the empirical histograms.

Using this approach to assess the $P(\kappa)$ of a real sequence would first involve creating an empirical distribution of possible κ values through repeated random shuffling. Once a distribution of κ values has been generated, a log-normal fit can be performed, and the probability of the κ value of interest determined from that functional form. Based on initial work it appears only 50-100 random permutants are required to build a basis set, from which the log-normal distribution can be generated. This analysis, when performed on the disordered regions in the human proteome, shows that the likelihood of observing IDP sequences with their naturally occurring κ values by random chance is essentially zero. This suggests strong evolutionary pressure away from the statistically expected random prior distribution of charged residues. It is important to remember that the expected value here refers only to the expected value given a uniform background prior. There are many additional constraints which influence how a set of amino acids are distributed in a linear sequence, but as a zeroth order approximation this provides some statistical context of the observed κ value for a given sequence.

Most likely κ value across diagram-of-state space

Considering the results of the preceding section, for a sequence of some given length and composition we can calculate the statistically expected κ value. If this is done for all sequence compositions of a given length, we can fully explore the κ-to-composition space. Fig. 4.18 shows a 2D heat map of four different sequence lengths (40, 60, 80, 100) where we calculated the κ values for all possible sequence compositions. The color in this heat map reports on expected κ value. A number of features emerge from fig. 4.19. Firstly, for the vast majority of sequence compositions the expected κ value is between 0.17 and 0.23. This result is relatively insensitive to sequence length. Secondly, in the cases of very strong polyelectrolytes (i.e.,

FCR ¿0.5) the expected κ value increases to 0.3 - 0.4. Finally, although not shown here or in fig. 4.17, when NCPR > 0.9 (i.e., the top left and bottom right corners of the diagram-of-states) the log-normal fitting procedure breaks down in much the same way as it does when FCR < 0.05. This inability to obtain a good fit is a result of one specific class of the residues (positive, negative, or neutral) entirely dominating the sequence composition and causing a rapid drop in the total number of possible sequence permutants.

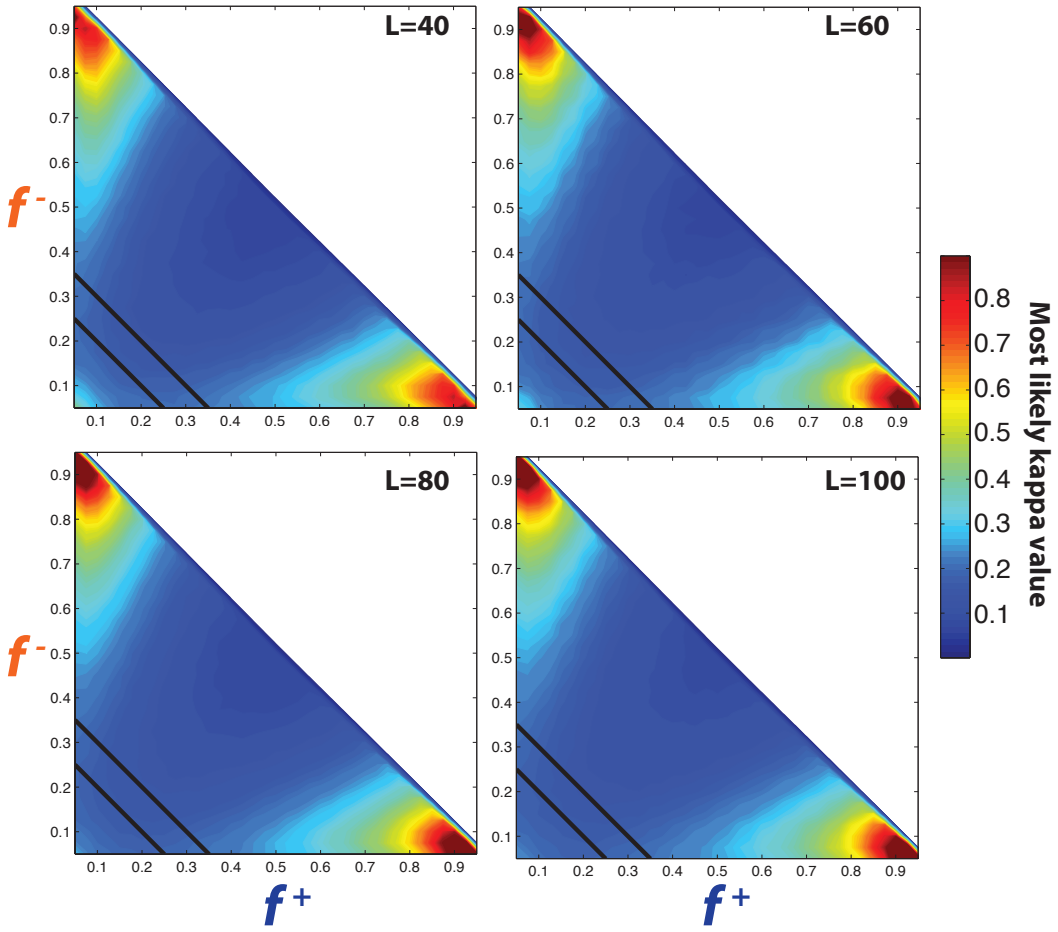

Figure 4.19: Statistically expected κ value given charge composition based on an unbiased uniform distribution of sequences. Expected values obtained by fitting a log-normal distribution to each sequence.

4.4 Discussion

In this section we discuss a small set of examples of how sequence analysis can be used to uncover inferences regarding the biophysical properties of IDPs. These inferences serve as ideal starting points for the development of testable hypotheses. In chapter 11 we used localCIDER to identify local clusters of high negative charge in the disordered region of the Nephrin intracellular domain (NICD) that drives phase separation via complex coacervation. The tools within localCIDER were coupled to a statistical analysis framework to identify the amino acid types that most strongly influenced phase separation based on an extensive mutagenesis screen. We identfieid tyrosine and arginine residues as crucial determinants of phase separation. Similarly, work by Nott *et al.* (which pre-dates the release of localCIDER) identified clusters of charged residues in the N-terminal IDR of Ddx4 [421]. These clusters are required to drive phase separation (see fig. 4.20). Given that many IDPs contain local clusters of charged residues, and that the patterning of charged residues has been shown to play a role in determining both conformation and function [30, 33, 125, 126], we expect there to be many more examples where the distribution of charged residues has a major impact on the sequence-ensemble-function relationships of IDPs.

Amino acid compositions of IDPs play central roles in determining their conformational properties [127, 359, 364, 405, 603, 666]. localCIDER enables rapid, proteome-wide investigations of compositional biases and the evolutionary preferences within IDPs. We analyzed the complete set of IDPs from sixteen model organisms to ask how general compositional biases in IDPs vary across diverse proteomes. For the higher eukaryotes (chordates), we found highly similar sequence properties, while lower eukaryotes displayed greater variety. The disordered proteome of *D. discoideum* showed a substantial deficiency of charged residues and enrichment in polar residues (notably Asn and Gln) when compared to other species. This

result is in accord with the findings of Malinovska *et al.* [356]. Conversely, the disordered proteome of *P. falciparum* is enriched in strong polyampholytic IDPs, with almost 50% of IDPs falling into R3 on the diagram-of-states. The complete analysis of 203,944 disordered fragments took just over two hours on a desktop computer, showcasing the high-throughput nature of localCIDER.

Fig. 4.20 shows three examples of the types of linear sequence analysis that localCIDER facilitates. In fig. 4.20a, the charge patterning associated with the N-terminal IDR from Ddx4 identified by Nott *et al.* is re-examined [420]. Fig. 4.20b illustrates the complexity and composition associated with the protein FUS, which is known to drive liquid-liquid phase separation *in vitro* and *in vivo* [443]. In addition to the well-characterized N-terminal low complexity domain (LCD), which we refer to as LCD1, we highlight two shorter LCDs towards the C-terminus (LCD2 and LCD3). To the best of our knowledge these regions remain largely unexplored and it is conceivable that they contribute to modulating the driving forces for phase separation. Finally, fig. 4.20c illustrates the charge distribution across the tau protein (4R-441 isoform). The N-terminal 120 residues encompass a high density of acidic residues, while the remainder of the sequence is highly basic. The delineation of positively charged and negatively charged residues does not overlap with other known sequence annotations. This charge distribution is expected to have an impact on how tau associates with other charged biopolymers such as heparin due to an effective macro-dipole across the sequence [203].

These results represent the tip of the iceberg. While not included in this discussion, in multiple ongoing projects we have used these sequence analysis tools coupled with additional approaches to understand, predict, redesign, and explore the sequence determinants of conformational behaviour. It has provided us with a powerful analytical framework, through

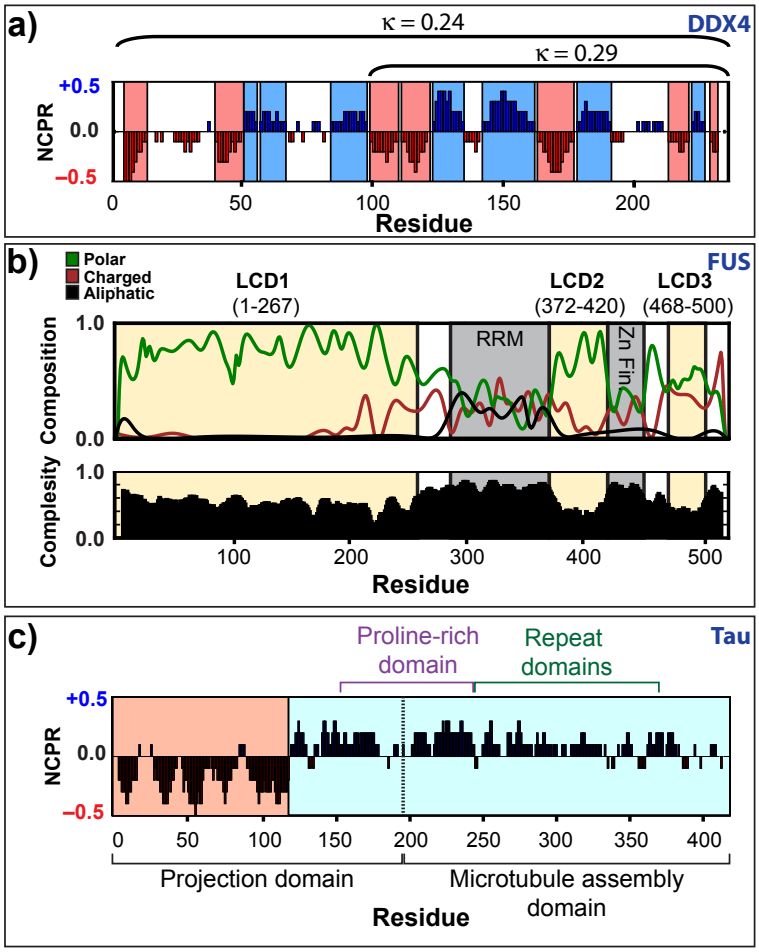

Figure 4.20: Three examples of linear sequence analysis performed by localCIDER. Panel (a) shows the charge patterning in the protein Ddx4, identifying local region with a high net positive or negative charge and showing the full sequence and C-terminal κ values. Panel (b) illustrates the local sequence composition and complexity of the protein FUS. RRMs and Zn-finger domains annotated based on published structural information. Panel (c) shows the charge distribution in the 4R-441 isoform, with various domains and regions annotated.

which we are converging on a high resolution understanding on how amino acid sequence determines conformational behaviour, and how that conformational behaviour determines function.

Chapter 5

A Dissection of Backbone and Sidechain Interactions

The following section is taken from the paper **Quantitative assessments of the distinct contributions of polypeptide backbone amides versus side chain groups to chain expansion via chemical denaturation** by A.S. Holehouse, K. Garai, N. Lyle, A. Vitalis, and R.V. Pappu. This was published in the *Journal of the American Chemical Society*, Vol. 137, pages 2984 - 2995, in February 2015. The text has been expanded to include additional detail. FCS experiments were performed by K.G. simulations were performed by A.S.H., A.V and N.L. All analysis was performed by A.S.H.

5.1 Background

Tanford's classical studies on protein structure and function established that functional activity and structural features of globular proteins are abrogated in the presence of high concentrations of denaturants such as 8 M urea and 6 M GdmCl. As introduced towards

the end of chapter 2, for unfolded proteins, polymeric properties that describe the ensemble average dimensions such as the hydrodynamic radius (R_H), radius of gyration (R_G) or end-to-end distance (R_{EE}) show a power law behaviour consistent with a polymer in a good solvent [572, 573]. This behaviour is captured by the relationship $R_H \propto N^{\nu}$, where N is the number of amino acids and ν a characteristic scaling exponent. Tanford showed that $R_H \propto N^{0.59}$ for highly denatured proteins [180]. Wilkins *et al.* used pulse-field gradient NMR measurements to recapitulate the scaling of R_H with N for a set of single domain proteins that show apparent two-state behaviour [641]. Similarly, Kohn *et al.* used SAXS to show that the mean radius of gyration (R_G) scales as $N^{0.59}$ for 28 different chemically denatured proteins of different lengths and amino acid sequences [297].

The overall implications of the scaling of R_H and R_G with N are two-fold: First, solutions with high concentrations of denaturants akin to good solvents for generic protein sequences. Second, given that many proteins show apparent two-state behaviour, the conjecture that emerges is that generic unfolded proteins sample ensembles with similar statistical properties. This conjecture has received considerable scrutiny and several lines of investigation have established that a scaling exponent of 0.59 does not imply purely self-avoiding random-coil-like conformations for denatured state ensembles [38,219,239,264,376,377,599,661]. Instead, the exponent of 0.59 accommodates considerable sequence specificity in the conformational properties of denatured proteins, allowing long-range contacts and local structure to exist in spite of apparent good solvent behaviour.

Given the apparent universality enforced by high concentrations of denaturant despite enormous sequence and native-state structural variation, our work is motivated by the question of why aqueous solutions with high concentrations of denaturants should be good solvents for generic proteins? Studies based on the solute partitioning model, atomistic simulations

and experimental data have converged on a consensus that urea denatures proteins through preferential interactions with backbone and sidechain atoms [87, 88, 164, 165, 211, 241, 243, 299, 342, 397, 476, 561, 562]. Specifically, urea molecules accumulate preferentially around the carbonyl oxygen atoms of peptide group amides, and to different degrees around the aliphatic, aromatic, and polar sites of sidechains [19, 211, 212, 476]. The mechanisms for denaturation in solutions with high concentrations of GdmCl remain unresolved although insights are emerging from different types of experiments [355]. Lim *et al.* measured the ability of guanidinium ions to block acid- and base-catalysed hydrogen exchange of an alanine dipeptide in high concentrations of GdmCl [333]. Their results suggest an absence of direct interactions between guanidinium ions and the functional groups of backbone amides. Studies with other model compounds suggest that guanidinium ions interact favourably with aromatic groups and primary amides of sidechains [367, 368]. Simulations suggest that the strengths of ion pairs are reduced in high concentrations of GdmCl [429]. These results highlight a prominent role for sidechain-mediated interactions as drivers of the loss of structure and chain expansion in solutions with high concentrations of GdmCl. The recent work of Jha and Marqusee suggests that denaturation follows a two-stage mechanism [265]. The first step appears to involve accumulation of guanidinium ions near the protein surface and this is followed by penetration of water molecules to disrupt the hydrophobic core.

Studies based on simulations and fluorescence correlation spectroscopy (FCS) experiments have established that water is a poor solvent for polypeptide backbones [576, 591]. In poor solvents, quantities such as R_G and R_H scale as $N^{0.33}$ thus ensuring that the chain-solvent interface is minimized [180]. Similar behavior has been observed using a combination of simulations and experiments for intrinsically disordered polar tracts such as polyglutamine, glycine-serine block copolypeptides, and the Gln / Asn rich N-domain of Sup35 protein [116,404,591]. It may be surprising that Gln and Asn rich sequences drive compact globules, given their

classification as polar residues, their favourable free energy of solvation, and their prevalence on the surface of proteins [72]. All these results originates primarily from the relative context in which these residues are found. In the sequence context of aromatic and aliphatic sidechain-containing residues, Gln and Asn are *relatively* hydrophilic, and their presentation on the surface of a globular protein allows for the sequestration of hydrophobic residues into the interior. In a polar rich sequence (such as polyQ or the Sup35 N domain), a self-solvation driven by extensive and degenerate backbone-sidechain and sidechain-sidechain, and backbone-backbone interaction drives collapse [116, 404]. These results reflect only part of the impact of sidechains on the conformational behaviour of polypeptides. For IDPs, the amino acid composition can drive an ensembles conformational behaviour between totally collapsed (e.g. polyQ) and highly expanded (protamine sequences) [359, 364, 405]. Importantly, while there are certain amino acids that are associated with different types of behaviour, the conformational behaviour is ultimate an emergent property of the composition and patterning of the sequence. These topics are dealt with extensively in chapters 2 and 4.

The preceding observations raise two questions that form the focus of our work:

1. Do polypeptide backbones, in the *absence* of sidechains, expand in a manner that is consistent with the observed scaling exponent of 0.59 in aqueous solutions with high concentrations of denaturants?

2. What role do sidechains play in influencing the expansion of polypeptide backbones in aqueous solutions with high concentrations of denaturants?

Answers to these questions provide deeper insights into the mechanisms of protein denaturation, and more generally provide a useful framework for thinking about protein-solute interactions. Our findings highlight the need to go beyond inferences gleaned from the

studies of model compounds. This is important if we are to obtain a coherent and comprehensive understanding of protein denaturation and the conformational properties of proteins in complex milieus such as cellular environments. The objects of our study are polyglycine peptides that mimic pure polypeptide backbones and two 15-residue peptides that serve as model systems to help elucidate the role of sidechains.

We report results from atomistic simulations and FCS experiments. The analysis of our simulation results is guided by the use of reference ensembles that mimic the conformational statistics of flexible polymers in poor, indifferent (Θ), and good solvents. We also introduce the effective concentration of backbone amides as a parameter to help in quantifying how backbone conformations are altered by the combination of sidechain-mediated interactions and preferential interactions of different sidechain groups with denaturants.

5.2 Methods

5.2.1 Peptide Systems

We used molecular dynamics (MD) simulations based on atomistic models for peptides and explicit representations of solvent and cosolute molecules to simulate the effects of water, 8 m urea and 8 m GdmCl on three different peptide systems. In order to assess the impact of denaturants on the conformational properties of pure polypeptide backbones, we performed three sets of simulations for a polyglycine peptide, N-acetyl-$(Gly)_{15}$-N-methylamide referred to hereafter as G_{15}. To understand how sidechains modulate the intrinsic properties of backbones in different environments, we performed simulations for two archetypal peptides designated as CAP **QFHFHWNRQDDQYFE** and OSP **GVSLLTIDVKKSLTK**.

The N- and C-termini were capped using N-acetyl and N-methylamide groups, respectively. These 15-residue peptides are based on fragments of full-length proteins and are excised from Carbonic Anhydrase (CAP) and from OspA (OSP). The sequences of CAP and OSP show negligible biases toward specific secondary or tertiary structures in water and they serve as useful model systems for unfolded states under folding conditions. The sequences have complementary attributes. CAP has no aliphatic residues whereas OSP has no aromatic residues. The net charge per residue is -0.2 for CAP and +0.2 for OSP. The fraction of charged residues is 0.27 for both peptides. Based on the combination of hydrophobicity, net charge per residue, and fraction of charged residues, these sequences and longer tandem repeats of these sequences are expected to have a predominant preference for heterogeneous distributions of globular conformations in water. Fluorescence correlation spectroscopy (FCS) experiments were performed for three peptides containing polyglycine tracts of different lengths. The peptides were of the form: Trp-$(Gly)_N$-Cys*-$(Lys)_2$ with $N = 15$, 31, and 45. Here, Cys* denotes a cysteine that was modified by covalent addition of an Alexa488 dye through a maleimide linkage. The Lys residues were necessary to enhance solubility and enable peptide purification and the Trp residue was used for accurate assessments of peptide concentration.

5.2.2 Molecular Mechanics Forcefields

We used the TIP3P model for water molecules [273]. We also used explicit representations for urea molecules and guanidinium (Gdm^+) and chloride (Cl^-) ions. We used the Kirkwood-Buff forcefield (KBFF) to model urea and GdmCl [634–636]. Molecular mechanics parameters for the three peptides and neutralizing counterions were taken from the OPLS-AA/L forcefield [276]. Neutralizing Na^+ and Cl^- ions were included in the simulations of CAP and OSP,

respectively. Our choices maintain fidelity with the paradigm for the development of the KBFF forcefield, which has been designed for interoperability with the OPLS-AA/L forcefield for peptides and neutralizing counterions. Recent work has highlighted issues with the combination of the OPLS-AA/L forcefield and the TIP3P water model for modelling conformational equilibria of various peptide systems [43, 44]. In this context, it is noteworthy that the collapse and poor solubility of polyglycine in water have been reproduced using other combinations of forcefields and water models, thus pointing to the robustness of the results to differences in forcefields.

5.2.3 Details of the Molecular Dynamics Simulations

We used version 4.5.3 of the GROMACS modelling package for the MD simulations [462]. The design of these simulations was based on the multiple-replica MD or MRMD approach of Vitalis *et al.* [616]. In this approach, one performs multiple independent simulations, each starting from an entirely different conformation for the peptide in question. The starting conformations are drawn at random from pre-equilibrated ensembles of sterically allowed conformations that are expanded and collapsed. Each simulation was designed to be long enough to ensure multiple recurrent transitions between compact globular conformations and expanded coil-like conformations. In high concentrations of denaturants, the increased viscosities slow the overall transitions. These considerations were used to set the upper limit on the simulation time for each replica. The parameters of the MRMD protocol were as follows. For each peptide in water and 8 m urea, each independent MD simulation was run for 110 ns and for these peptides in 8 m GdmCl the simulation time per replica was 210 ns. For each of the replicates, the first ten nanoseconds of simulations were set aside as equilibration. Overall, for each combination of peptide and environment we performed

20 independent simulations. This yielded an aggregate simulation time of 2.1 µs for each of polyglycine, CAP, and OSP in water and 8 m urea and an aggregate simulation time of 4.1 µs for each of the three peptides in 8 m GdmCl.

The equations of motion were integrated using the leapfrog integrator with a timestep of 2 fs. All peptide bond lengths and those within urea molecules and Gdm^+ ions were constrained using the LINCS algorithm [230]. The bonds and angles within TIP3P water molecules were constrained using the SETTLE algorithm [396]. The simulations were performed in the isothermal-isobaric ensemble. The target temperature, pressure, and isothermal compressibility in all simulations were 298 K, 1 bar, and 4.5×10^{-5} bar^{-1}, respectively. We used the velocity rescaling method of Bussi *et al.* with a coupling constant of 1.0 ps to control the temperature [81]. The simulation pressure was controlled using the extended-ensemble barostat of Parrinello and Rahman [442]. The coupling time for the latter was 20 ps. Snapshots were saved once every 12.5 ps. Each snapshot included the positions of the peptide atoms and those of the denaturant molecules (urea and Gdm^+ and Cl^- ions).

In each of the MRMD simulations we used cubic boxes with periodic boundary conditions. Long-range electrostatic interactions between periodic images were treated using the particle mesh Ewald approach [123]. We used an eighth-order cubic interpolation with a tolerance of 10^{-5}. Cutoffs of 11 Å and 14 Å were used for the real space electrostatic and van der Waals interactions, respectively. Long-range dispersion corrections were applied for energy and pressure. Neighbour lists were updated once every five steps. This choice ensured against large deviations from the target pressures in all of the MD simulations. The average dimensions of the box as prescribed by the average length to each side ranged from 61 Å for peptides in water to 71 Å for peptides in 8 m GdmCl. The maximum end-to-end distance of each peptide is ca. 60 Å and this value is never realized even in ensembles of

self-avoiding random walks. Hence, the dimensions of the central simulation cell were large enough to accommodate maximally extended conformations and rule out any compaction due to artefacts imposed by confinement. In all of the simulations, we fixed the number of water molecules to be 7,360. For simulations in 8 m urea, we used 1,060 urea molecules and for simulations in 8 m GdmCl we used 1,060 Gdm^+ and 1,060 Cl^- ions. The choice for the number of water molecules was made to ensure a density of 1 gm/cm^3 in a periodic box of volume 2.16105 Å^3. In denaturing environments, the density of water is maintained by the increase in the box size, which is necessary to accommodate denaturant molecules. We used molality rather than molarity because molality remains constant irrespective of volume fluctuations.

5.2.4 Simulations & Analysis of Reference Ensembles

For each peptide, we generated reference ensembles using potentials that encode conformational properties corresponding to three distinct model scenarios. For these simulations we used version 1.0 of the CAMPARI modelling package (http://campari.sourceforge.net). For each peptide, we performed two sets of reference simulations using the ABSINTH model while zeroing out the mean field solvation and Coulomb terms of the potential. All other terms of the potential were used as prescribed by the ABSINTH model [613]. The two reference potentials are distinguished by the choice of in equation 5.1. In one set of reference simulations, $\lambda = 0$ and in the other $\lambda = 1$.

$$U_{ref} = 4 \sum_i \sum_{j<i} \epsilon_{i,j} \left[\left(\frac{\sigma_{i,j}}{r_{i,j}} \right)^{12} - \lambda \left(\frac{\sigma_{i,j}}{r_{i,j}} \right)^{6} \right] \tag{5.1}$$

The summation runs over all unique pairs of non-bonded atoms as defined by the ABSINTH model [613]. Metropolis Monte Carlo simulations were performed at a simulation temperature of 298 K. The parameters for ij, ij and other non-zero terms of the potential were taken from the abs3.2_opls.prm parameter file that is part of the CAMPARI distribution. To generate an ensemble consistent with a good solvents (referred to hereafter as the excluded volume [EV] ensemble) λ is set to 0 in equation 5.1. This has the effect of turning off all attractive interactions, leaving only repulsive Lennard-Jones interactions. To generate an ensemble consistent with a poor solvent (referred to hereafter as the Lennard-Jones (LJ) ensemble] λ is set to 1 in equation 5.1. This has the effect of converting the polypeptide into an approximately uniformly sticky polymer with no-long-range repulsions, but where local steric excluded volume effects are respected. Quantities such as R_G and R_H scale as $N^{0.33}$ as a function of chain length for all systems in the poor solvent limit limit and as $N^{0.59}$ in the good solvent limit. R_{ee} can show similar scaling, but the R_{ee} scaling can become uncoupled for compact polymers [206].

We also performed reference simulations using the rotational isomeric approximation to mimic the Flory random coil (FRC) limit. To generate the FRC limit a pre-generated database of locally allowed residue conformations was generated. This was done using the ABSINTH model with $\lambda = 0$ in equation (1) combined with the mean field solvation and electrostatic terms being zeroed out. Dipeptides (i.e., Ac-Xaa-Nme) simulations for all twenty amino acids were performed using Metropolis Monte Carlo at 298 K. The distributions of ϕ, ψ, and χ angles from the dipeptide simulations were used to create libraries of rotational isomers for each amino acid. To generate FRC ensembles for longer chains ϕ, ψ, and χ angles were randomly drawn from the residue-specific libraries of rotational isomers. In these simulations all inter-residue interactions between are explicitly zeroed out. The resultant ensembles conform to Flory's approach for mimicking conformational distributions that result

from the counterbalancing of chain-chain and chain-solvent interactions in an indifferent or theta solvent. Quantities such as R_G, R_{ee}, and R_H scale as $N^{0.5}$ as a function of chain length for all systems in the FRC limit. Similarly, distributions for a range of polymeric quantities match expectations from theory and simulation for chains in a theta solvent [181, 559].

5.2.5 Parameters that Quantify Chain Size and Shape

In a given environment, for each snapshot, we calculated the gyration tensor defined as:

$$\mathbf{T} = \frac{1}{n_a} \sum_{i=1}^{n_a} (r_i - r_c) \otimes (r_c - r_i) \tag{5.2}$$

Here, r_i is the position vector of atom i within a specific conformation, r_c is the location of the centroid for this conformation, n_a is the number of atoms in the chain, and the symbol $\otimes$ refers to the dyadic product. We use the eigenvalues L_j $(j = 1, 2, 3)$ of the gyration tensor for the specific conformation to calculate two global descriptors of conformations; the radius of gyration (R_G) and the asphericity (δ^*) [559].

$$R_G = \sqrt{L_1 + L_2 + L_3} \tag{5.3}$$

$$\delta^* = 1 - 3\frac{L_1L_2 + L_2L_2 + L_3L_1}{(L_1 + L_2 + L_3)^2} \tag{5.4}$$

R_G is a formal order parameter in polymer theories and serves as a measure of chain density. The δ^* is a measure of the shape associated with a particular conformation. The values

of $\langle\delta^*\rangle$ are predicted by theory to be approximately 0.42 and 0.52, for long, linear, flexible chains in theta (FRC limit) and good solvents (EV limit), respectively whereas $\langle\delta^*\rangle \to 0$ for compact globules. For globules formed by finite sized linear chains, δ^* ranges between 0.05 and 0.3, with the smaller values corresponding to longer chains.

5.2.6 Calculation of Internal Scaling Profiles

We utilized internal scaling profiles to compare the ensemble-averaged conformational properties of polypeptide backbones for different systems in different milieus [360]. For a specific linear sequence separation $|i-j|$, we calculated $R|i-j|$ as follows:

$$\langle\langle R\rangle\rangle_{|i-j|} = \left\langle \frac{1}{Z_{ij}} \sum_{m\in i} \sum_{n\in j} |r_m^i - r_n^j| \right\rangle \tag{5.5}$$

r_m^i and r_n^j are the position vectors of backbone atoms m and n from residues i and j, respectively. Z_{ij} is the number of unique pairwise distances between the backbone units of residues i and j. The internal scaling profiles describes relationship between $\langle\langle R_{|i-j|}\rangle\rangle$ and $|i-j|$. This provides a robust classifier of conformational behaviour though a complete albeit highly averaged description of the conformational properties across all length scales [360]. The notation for$\langle\langle R_{|i-j|}\rangle\rangle$ is intended to clarify the fact that the averaging is over all conformations in the ensemble (the outer average) for all pairs of residues that are $|i-j|$ apart in the linear sequence (the inner average).

5.2.7 Sample Preparation for FCS Measurements

Peptides of $WG_{15}CKK$, $WG_{31}CKK$ and $WG_{45}CKK$ were purchased in crude form from Yale University's Keck peptide synthesis facility. The identities of the peptides were confirmed using mass spectrometry. For each peptide, the powder was suspended in water at 1 mg/ml concentration. The suspension was then sonicated for two minutes using a tabletop water bath sonicator. Since polyglycine is practically insoluble in water, LiCl powder (1 mg/ml) was added to this solution and dissolved by vortexing to obtain a clear solution. Tris(2-carboxyethyl)phosphine (TCEP) at 1 mM concentration was added to the solution to reduce any pre-formed disulphide bonds. The pH was adjusted to 7.4 using a 20 mM Hepes buffer. Finally, 200 µM Alexa488 maleimide dye was added, and the solution was incubated at room temperature for 3 hrs. This solution was then stored overnight at 4°C. Free dyes were removed by dialysis of the solution for 24 hrs in water in the presence of 5 mM β-mercaptoethanol using a 2 kDa dialysis membrane (Spectrapor). Centrifuging the sample and discarding the supernatant removed any free dye that remained following dialysis. The pellet containing the labelled polyglycine peptide was dissolved in an aqueous solution of 8 M LiCl. The peptide was further purified by size exclusion chromatography using a superdex peptide column (GE healthcare). The labelling efficiency, determined by the absorbance of the peptide at 488 nm and 280 nm, was found to be $> 80\%$ in all cases. The concentrations of purified and labelled peptides in the final stock solutions were 6, 4 and3 µM for $WG_{15}C^*KK$, $WG_{31}C^*KK$ and $WG_{45}C^*KK$ respectively.

5.2.8 Details of FCS Measurements

FCS has been used to reproduce the dimensions of highly expanded systems in the presence and absence of denaturants [359,539]. Here, we used a Zeiss confocor 2 microscope equipped with FCS measurement capability. For the diffusion measurements, the stock solutions of Alexa488-labelled polyglycine peptides were diluted by 100-fold into water, urea (4 M and 8 M) or GdmCl (3.5 M and 7 M). The measured diffusion times were found to be insensitive to further dilution. The measurements were also performed on a free Alexa488 dye (50 nM) solution in each of the solvent conditions as controls. Measurements in each condition were done in triplicate. In order to avoid optical aberrations due to high refractive indices in urea and GdmCl solutions, all of the measurements were performed at depths within 4-6 µm from the cover glass. The FCS auto-correlation traces were fit using one triplet and one diffusing species. To calculate the intrinsic diffusion time, we calculated a correction factor, which we defined as the observed diffusion time for the free dye in water divided by the diffusion time for the free dye in the environment of interest. Since the dye does not undergo any change in conformation under denaturing conditions, the multiplicative correction factor provides a route to generate environment-corrected values, which we designate as the intrinsic diffusion time for the peptide in the environment of interest.

Water is a poor solvent for polypeptide backbones. In poor solvents, there exists a saturation concentration beyond which the polymer plus solvent system separates into solvent-rich and insoluble polymer-rich phases [472, 473]. Polyglycine and polyglutamine are examples of polypeptide polyamides. The measured saturation concentrations for a range of polyglutamine peptides of different lengths are in the low- to sub-micromolar range and these saturation concentrations decrease with increasing polyglutamine length [117]. Below the

saturation concentration, there exists a second saturation boundary that is akin to a micellization boundary where the critical micelle concentration is ∼100 nM or lower [117]. The data for polyglutamine and observations for glycine-rich peptides are consistent with our findings that polyglycine peptides are highly insoluble in water [18,576]. This should in turn yield globules for individual chains in ultra dilute solutions for polyglycine in water [472,473,576]. From a practical standpoint, measured saturation concentrations place constraints on the concentration ranges one can use for measuring the conformational properties of individual polypeptides. Measurements of hydrodynamic properties have to be performed in the low nanomolar or even picomolar concentrations, depending on chain length. According to the Flory theorem, an individual chain within an aggregate can have dimensions that scale as $N^{0.5}$ if the aggregates are reasonably large. This taken together with the lower diffusivity of aggregates will confound interpretations of measured diffusion times. Our data were collected at concentrations that lie below the known / inferred saturation concentrations and critical micelle concentrations for polypeptide polyamides. Further, the brightness per molecule matches that of the free dye implying the absence of aggregates and the monomeric form being the only diffusing species in all experiments.

5.3 Results

Our overall approach is to obtain the conformational distributions for the polypeptide backbones of polyglycine, CAP, and OSP in water, 8 m urea, and 8 m GdmCl and compare these to distributions obtained for the same systems modelled in the LJ, FRC, and EV limits.

5.3.1 Quantifying Impact of Denaturant on Polyglycine

Fig. 5.1 shows the mean values for R_G and δ^* that were obtained for G_{15} in water, 8 m GdmCl, 8 m urea, and the three reference ensembles, respectively. The mean R_G and δ^* values suggest a systematic expansion of G_{15} in the two denaturing environments. The degree of expansion is higher in urea than GdmCl, although the degrees of expansion observed in both denaturing environments is significantly less than would be expected if the polypeptide were undergoing true denaturation. Notably, the degree of expansion in the denaturing environments is smaller than in the FRC or EV reference ensembles.

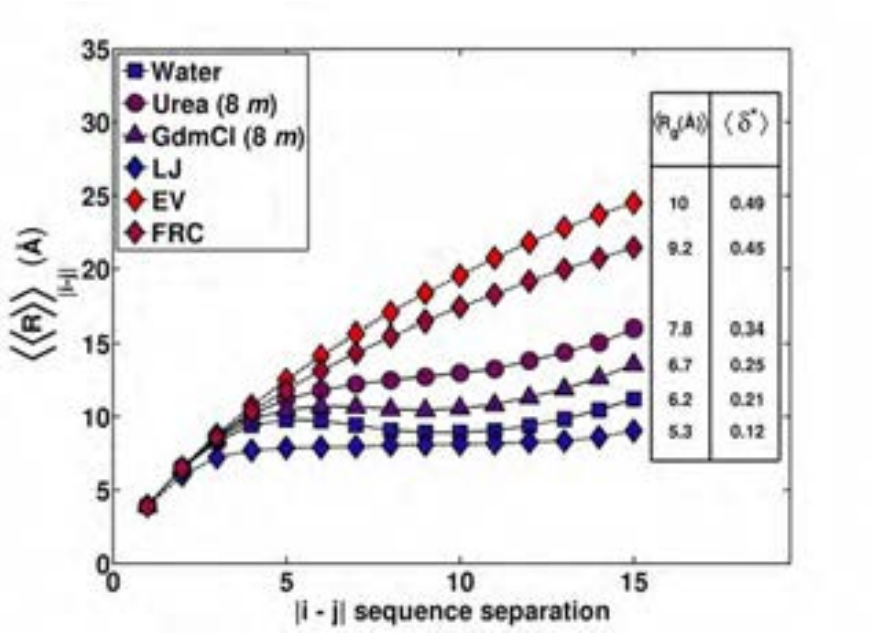

Figure 5.1: Internal scaling profiles for G_{15} in water, 8 m urea, 8 m GdmCl compared to similar profiles calculated for G_{15} in the EV, FRC, and LJ limits. Error bars are excluded in the interest of clarity. The legend shows the mean R_G and δ^* values for the three environments and the three reference ensembles.

A distinct feature of internal scaling profiles for the FRC and EV reference ensembles is the monotonic increase of $R_{|i-j|}$ with linear sequence separation $|i-j|$. This behaviour derives from the fractal nature of flexible chains in the FRC and EV limits. In contrast, the profile for the LJ reference shows plateauing behaviour, and the densities of the globules that form will dictate the plateau values. fig. 5.1 shows that the profiles for G_{15} in 8 m urea and 8 GdmCl exhibit signatures of this plateauing behavior, suggesting a persistent preferences for globular conformations as observed for polyglycine in water. The plateau values obtained in denaturing environments are larger in denaturant than they are in water, as are the mean radii of gyration. Interestingly, despite conventional wisdom that GdmCl is a stronger denaturant that urea, urea invokes a stronger response for G_{15} than GdmCl does - the difference between GdmCl and urea is greater than the difference between GdmCl and water.

Do the internal scaling profiles imply uniformly swollen globules in 8 m urea and 8 m GdmCl or do they imply increased sampling of expanded conformations via spontaneous fluctuations? To answer this question we performed a comparative analysis of the joint distributions $P(\delta^*, R_G)$ calculated for G_{15} in each of the three environments and each of the three reference ensembles. These distributions are shown in Fig. 5.2. We quantify the populations for distinct asphericity intervals to compare the amplitudes of conformational fluctuations in different milieus.

The fluctuations in sizes and shapes are correlated, and this diminishes the possibility of sampling conformations with high R_G and low asphericity values, thus ruling out uniformly swollen globules in denaturing environments. Instead, the ensembles in 8 m urea and GdmCl are mixtures of compact spherical conformations and expanded aspherical conformations. In 8 m urea there is a 30% reduction in the population of compact spherical conformations

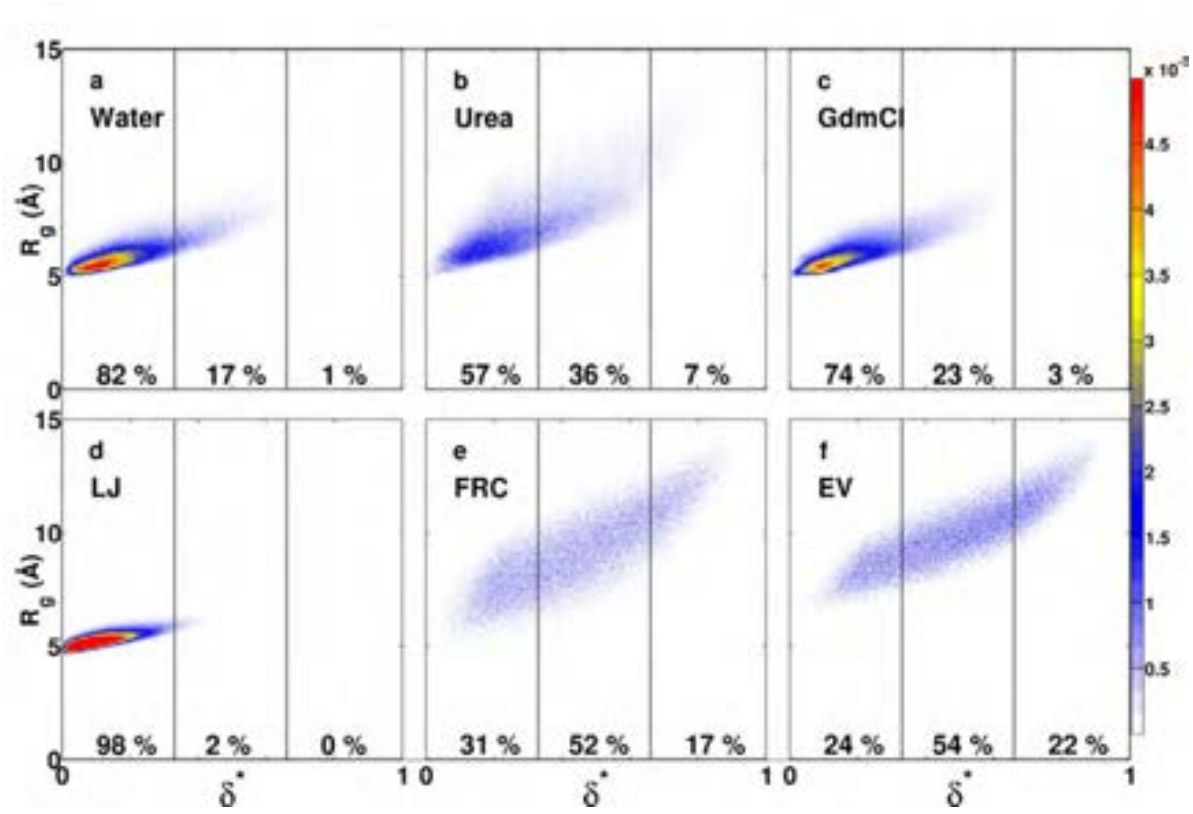

Figure 5.2: Plots of the joint probability densities $P(R_G, \delta^*)$ of sizes and shapes for G_{15} in water, 8 m urea, and 8 m GdmCl (top row) and in the LJ, FRC, and EV limits (bottom row). Each panel also shows the populations within three distinct, equally sized, non-overlapping intervals along the δ^* axis.

compared to the population in water, and this population is reduced by 10% in 8 m GdmCl. However, in order to achieve statistics that are congruent with those of canonical random coils such as the FRC or EV reference ensembles, the population of compact spherical conformations has to be reduced by at least 60%. Clearly, this degree of expansion is not achieved for polypeptide backbones in high concentrations of urea and GdmCl and there remains a persistent preference for compact globular conformations.

5.3.2 Experimental Tests of Simulation Results

Fig. 5.3 summarizes results from FCS measurements for three polyglycine peptides in water at different concentrations of urea and GdmCl. In a given environment, the intrinsic diffusion times (τ_D) increase with chain length. Further, for a given chain length, the values of τ_D are highest in 4 and 8 M urea, respectively. In 3.5 M GdmCl the values of τ_D are similar to those in water and there is a small increase of τ_D in 7.5 M GdmCl. These results, shown in panel a of fig. 5.3, imply a higher degree of expansion for polyglycine chains in higher concentrations of urea as opposed to GdmCl. The value of τ_D measures the mean diffusion time through the confocal volume and this quantity is proportional to R_H.

Is the expansion we observe in denaturants congruent with expectations for chains in either the FRC or EV limits? We answer this question by performing a scaling analysis using the measured τ_D values for different chain lengths in different milieus. Given that $\tau_D \propto R_H$ it follows that $\tau_D \sim \tau_0(Mw)^\nu$ where Mw refers to the molecular weight of the diffusing species (including the dye) [116]. For each combination of peptide and environment, we obtained three independent estimates for D, plus a separate estimate for D of the free dye. Therefore, for a given milieu, we used multiple combinations of independent estimates of D to generate synthetic datasets for linear regression analysis of $\ln(\tau_D)$ as a function of $\ln(Mw)$. Each synthetic dataset has four data points, three for the labelled peptides and one for the free dye. The results do not change materially if we exclude the free dye from this analysis. For each of the five environments, we apply the following procedure to estimate the scaling exponent for polyglycine in that environment: (i) we randomly selected a set of four D values from the data replicates for the dye and the three peptides. (ii) We perform linear regression analysis by plotting $\ln(\tau_D)$ against $\ln(Mw)$. The slope of the line of best fit is an estimate of for the particular combination of four data points. For each regression attempt, the goodness

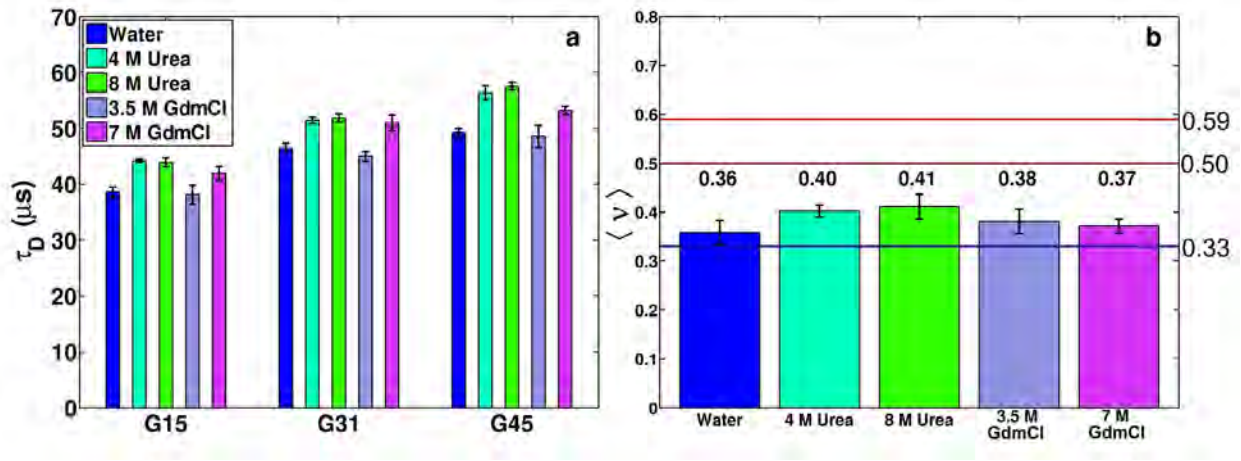

Figure 5.3: Summary of results from FCS experiments. Panel a shows the estimated values of τ_D in microseconds for three different polyglycine peptides in different milieus. Panel b shows the estimated scaling exponents extracted from the the scaling of τ_D as a function of molecular weight for polyglycine peptides in different milieus.

of fit was evaluated and on average, the regression lines were found to fit the data with no more than 1-2% overall error. (iii) Steps (i) and (ii) were repeated 1×10^4 times for each environment thereby yielding a distribution of 1×10^4 estimates for ν. These distributions were used to estimate the mean and standard deviation of ν for polyglycine in a specific solution environment.

The results of the scaling analysis are shown in panel b of Fig. 5.3 for polyglycine in water, 4 M urea, 8 M urea, 3.5 M GdmCl, and 7.5 M GdmCl, respectively. Our estimates for the values of for polyglycine in water, 4 M urea, 8 M urea, 3.5 M GdmCl, and 7.5 M GdmCl are 0.36 ± 0.03, 0.40 ± 0.01, 0.41 ± 0.03, 0.38 ± 0.03, and 0.37 ± 0.01, respectively. These results support the following conclusions: within bounds imposed by finite size artefacts, we can assert that water is a poor solvent for polyglycine. Further, although solvent quality improves in solutions with high concentrations of urea or GdmCl these milieus cannot be

classified as good solvents for polypeptide backbones. Taken together, the simulation results and assessments of experimental data yield mutually consistent inferences. Polypeptide backbones form compact globules in water and the despite discernible destabilization of the globules, the degree of expansion is insufficient to classify denaturing environments as good solvents for backbones. Instead, in denaturing environments, backbones sample a mixture of expanded and collapsed states, with a clear bias for the latter. It is worth reiterating that in both the simulations and experiments urea is a substantially more effective denaturant than GdmCl for the polypeptide backbone.

Taken together, our results suggest that the observed expansion of generic protein sequences in highly denaturing environments must derive mainly from the influences of amino acid sidechains [297]. Considering this, the question of interest now becomes through what mechanism to the sidechains engender interaction with the denaturant?

5.3.3 Sidechains Facilitate the Expansion of Polypeptide Backbones

Fig. 5.4 and 5.4 summarize results for two archetypal sidechain containing peptide sequences designated as CAP and OSP, respectively. With one exception, all residues in CAP and OSP are non-glycine residues, meaning backbone dihedral angles are limited by local steric hindrances that is not present for glycine. The EV, FRC, and LJ reference states account for this local steric hindrance.

Polyglycine is 17% more expanded in water than for the reference LJ globule. In contrast, the backbone is 25% more expanded in water for CAP and OSP as compared to the corresponding reference LJ globule. Therefore, sidechains can prime the backbone by inducing an

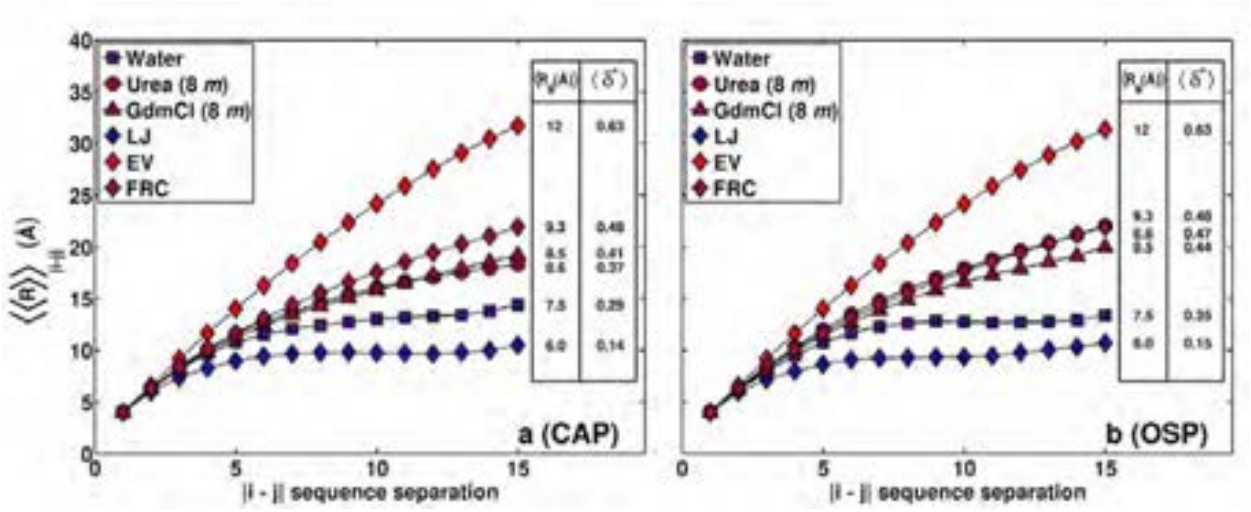

Figure 5.4: Internal scaling profiles for CAP and OSP

intrinsic expansion whereby the chain dimensions increase even in the absence of denaturant molecules.

The mean R_G and $\delta*$ values for the backbones of CAP and OSP in 8 m urea and 8 m GdmCl are closer to the FRC limit than is the case for polyglycine. These values are shown in Fig. 5.4 along with the internal scaling profiles, which provide visual evidence of the similarities between intra-backbone distances for the two peptides in the FRC limit and in denaturing environments. In order to enable direct comparisons to the results in Fig. 5.1, the internal scaling profiles shown in Fig. 5.4 were calculated using only backbone atoms. The sidechain priming of backbones is also illustrated by comparing the distributions for R_G and δ^* shown in panels a and g from fig. 5.5 to that of panel a in fig. 5.2. In water, there is a significant diminution in the population of compact spherical conformations and an increase in the population of more expanded aspherical conformations, especially for OSP, which has no residues with bulky aromatic sidechains. The distributions of R_G and δ^* values in 8 m urea and GdmCl show close agreement with those of the FRC limit, especially

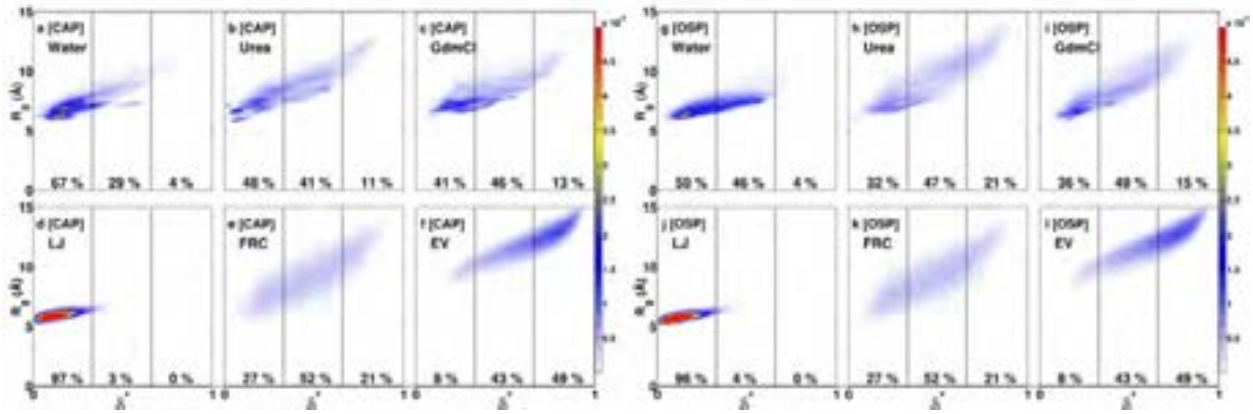

Figure 5.5: Distributions of R_G and δ^* values for the backbones of CAP and OSP in water, 8 m urea, and 8 m GdmCl (top row) compared to the equivalent distributions in the reference LJ, FRC, and EV ensembles (bottom row). Each panel shows the populations in three equally sized non-overlapping intervals along the δ^*-axis.

for the backbone of OSP. The increased expansion of OSP's backbone in water and in both denaturing environments is attributable to the lack of aromatic sidechains and to the presence of smaller aliphatic residues.

5.3.4 Quantifying the Convergence Toward Random Coil Ensembles

In Fig. 5.6 we quantify the effective concentrations of backbone amides for each of the three peptides in different environments and in the three reference ensembles. The values for the FRC and EV ensembles set the targets that are to be achieved for the effective concentrations if the ensembles are to converge upon one of the two canonical random coils. The effective concentration of amides is 19.2 M for polyglycine in water. This decreases to 17 M in 8 m GdmCl and 11.3 M in 8 m urea. However, the concentrations for polyglycine in the FRC

and EV ensembles are 6.7 M and 4.8 M, respectively. Despite a 41% dilution of the effective amide concentration caused by chain expansion in 8 m urea, the conformational properties of the backbone do not converge upon either of the random coil ensembles. In order to converge on the FRC limit, chain expansion needs to engender at least a 65% dilution of the effective amide concentration.

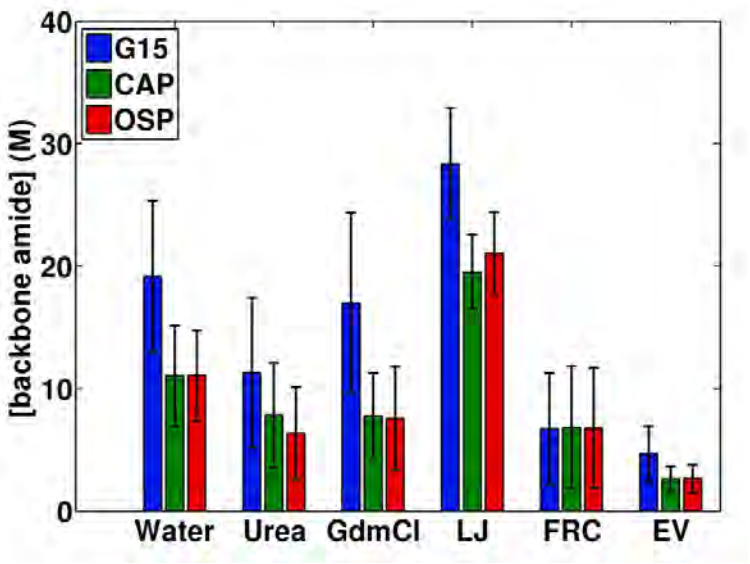

Figure 5.6: Effective concentrations of backbone amides and fluctuations calculated using the average R_G values and their standard deviations for G_{15}, CAP, and OSP.

To assess the impact of sidechains on the local concentration of backbone amides we calculated the effective amide concentration in water for CAP and OSP and compared the resulting value to G_{15}. For CAP and OSP the backbone amide concentration is ca. 11 M. Comparing this value with the local concentration for G_{15}, the sidechains act as a local solvent, inducing a 42% reduction in the effective amide concentrations for CAP and OSP vis--vis polyglycine in water. This reduction is similar to the extent of dilution realized by polyglycine in 8 m urea. For CAP and OSP the effective concentrations of backbone amides are ca. 6.7 M and 2.7 M, for the FRC and EV limits, respectively. Chain expansion induced

by denaturants leads to a further 39% dilution and fig. 5.6 shows that the concentrations for the FRC limit are achieved on average and as a result of conformational fluctuations for CAP and OSP in high concentrations of denaturants. In order to achieve congruence between the conformational properties of polypeptide backbones in denaturants and those of canonical random coils, there must be a suitable sidechain-mediated intrinsic expansion of the backbone in water in the absence of denaturants. We refer to this sidechain mediated expansion as *priming*.

5.3.5 Quantifying Solvent-Peptide Preferential Interactions

We used the integrals of site-site radial distribution functions to calculate the relative occupancies of denaturant molecules around different chemical moieties. These relative occupancies serve as proxies for preferential interaction coefficients that underlie the formalism of the solute partitioning model and analysis based on Kirkwood-Buff integrals [195, 299, 454, 499]. The relative occupancy parameters, denoted as π, were calculated using the following procedure: For a given combination of atomic sites denoted as X on the urea molecules and Y on a peptide we calculated:

$$\pi_{XY} = \frac{\int_{0\text{Å}}^{4\text{Å}} g(r_{XY})\; r_{XY}^2\; dr_{XY}}{\int_{0\text{Å}}^{4\text{Å}} g_{urea}(r_{NO})\; r_{NO}^2\; dr_{NO}} \tag{5.6}$$

Here, $g(r_{XY})$ is the radial distribution function that quantifies the relative probability of finding sites labelled X (either nitrogen or oxygen) on urea molecules within a distance r_{XY} around peptide sites of type Y. Similarly, $g_{urea}(r_{NO})$ is the radial distribution function that

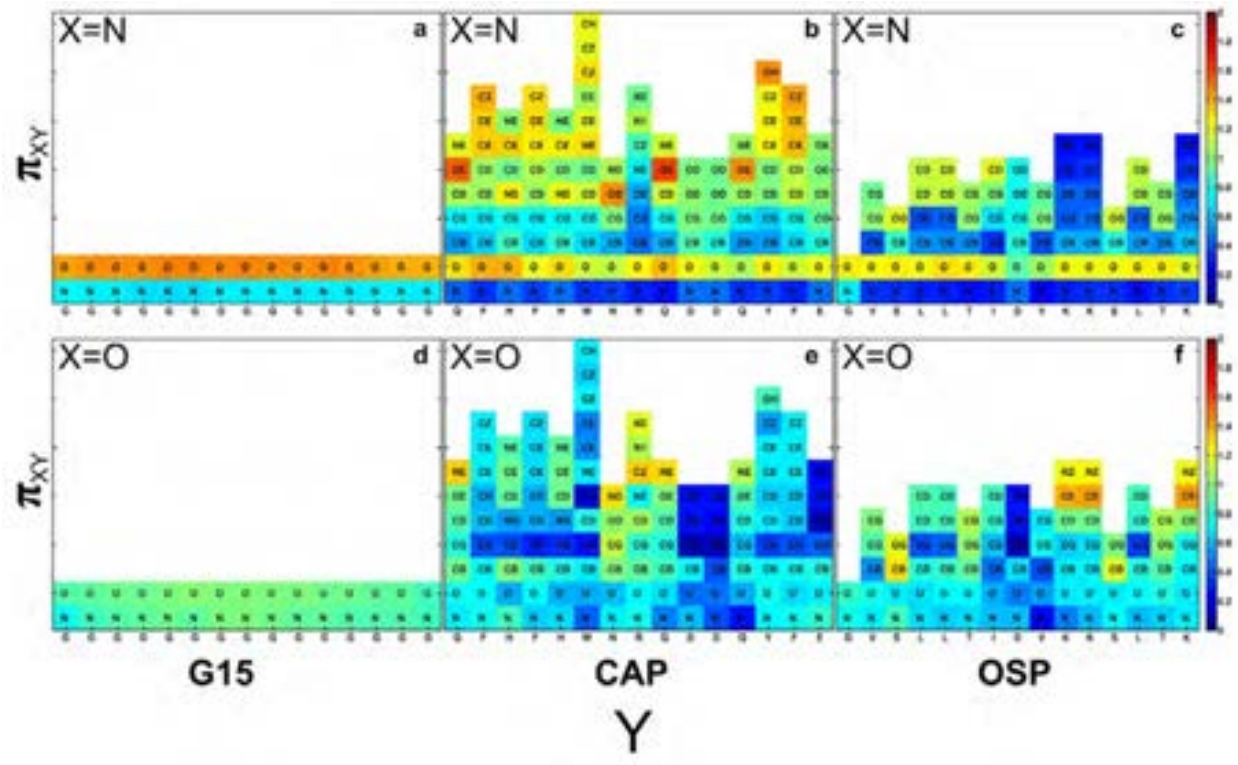

Figure 5.7: Values of π_{XY} for urea nitrogen (top row) and urea oxygen atoms (bottom row) around backbone and sidechain sites.

quantifies the relative probability of finding nitrogen atoms from urea molecules at a distance r_{NO} in the bulk solution from oxygen atoms on other urea molecules. We focus only on the effects of direct interatomic interactions including hydrogen bonds, and therefore we consider a length scale of 4 for each of the radial distribution functions. If π_{XY} is greater than unity, then there is accumulation of the urea site X around the peptide site Y and conversely, values of π_{XY} less than unity point to depletion of urea sites X around the peptide sites Y. The results obtained for peptides in 8 m urea are shown in Fig. 5.7.

The equivalent procedure was performed for GdmCl-peptide interaction using equation 5.7.

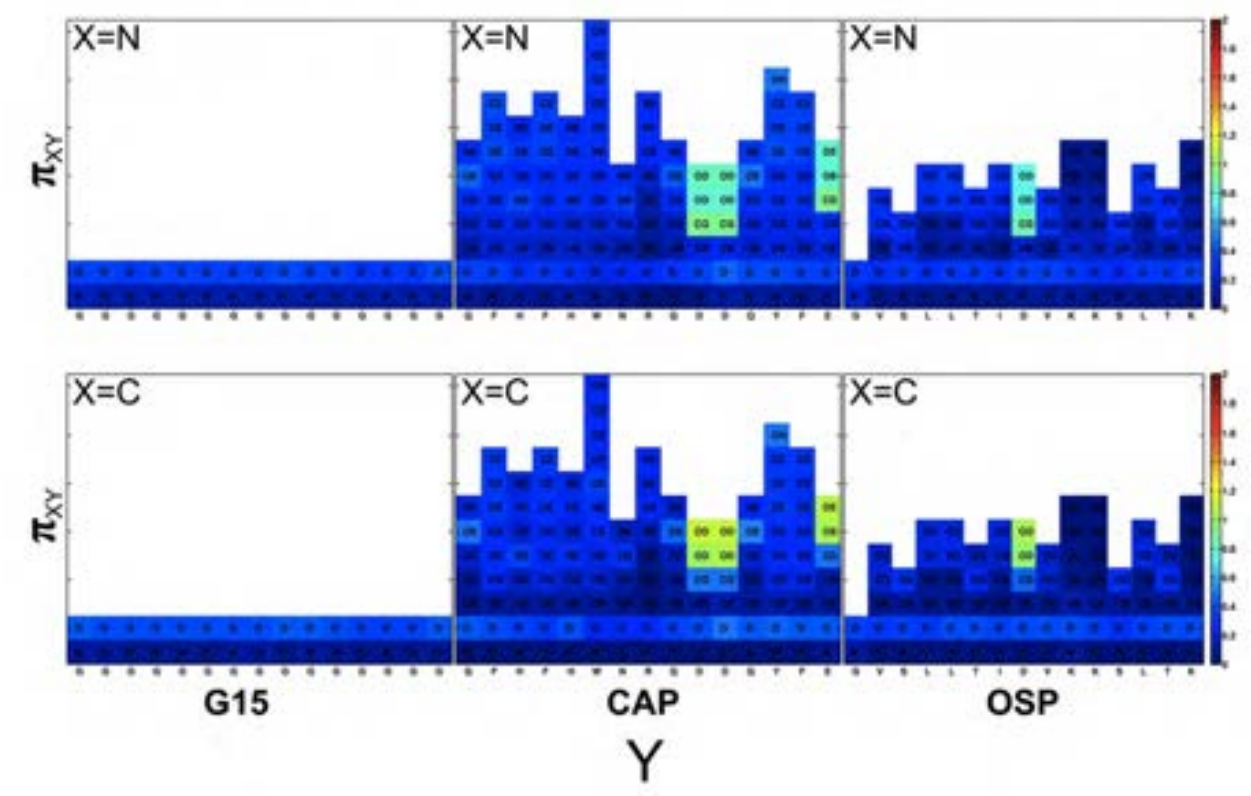

Figure 5.8: Relative occupancies of the Gdm^+ nitrogen atoms (top row) and central carbon atom around different backbone and sidechain sites of polyglycine, CAP, and OSP.

$$\pi_{XY} = \frac{\int_{0\text{Å}}^{4\text{Å}} g(r_{XY})\, r_{XY}^2\, dr_{XY}}{\int_{0\text{Å}}^{4\text{Å}} g_{GdmCl} u(r_{Gdm^+Cl^-})\, r_{Gdm^+Cl^-}^2\, dr_{Gdm^+Cl^-}} \tag{5.7}$$

The results for this analysis are shown in Fig. 5.8.

Our definition of π_{XY} is analogous, although not identical, to the definition of preferential interaction coefficients or partition coefficients that are central to the quantification of group-specific contributions to protein denaturation [138, 211, 476]. The central distinction is that unlike π_{XY}, which uses the strengths of donor-acceptor interactions between urea

molecules or interactions between Gdm^+ and Cl^- ions for GdmCl as the reference states, canonical preferential interaction / partition coefficients are referenced to interactions between urea / Gdm^+ with water molecules. Unfortunately, given the large box sizes, the numbers of independent simulations being performed, and our efforts to keep the storage demands tractable, we decided against saving the positions of water molecules for our simulations with denaturants. This choice, post facto, necessitated the use of a different reference state. Given the near ideality of urea-water mixtures our choice of reference state does not have a material impact on quantitative comparisons between our numbers for π_{XY} and those reported by Record and co-workers based on vapour pressure osmometry measurements for model compounds [138, 211, 212, 299, 476, 499, 635]. However, in GdmCl, additional complications are introduced by the favourable solvation of the Gdm^+ ion and electrostatic repulsions/attractions with other Gdm^+/Cl^- ions in the bulk solution. This confounds our analysis of the site-site pair correlations because the energy scales that contribute to the reference distributions are fundamentally different and hence the values of π do not lend themselves to ready interpretations regarding accumulation versus depletion. Although reasonable inferences can be gleaned from the relative trends of Gdm^+ occupancies around different sites, quantitative comparisons to experimental data will require the use as reference the pair correlation functions that quantify the strengths interactions between Gdm^+ and water molecules as opposed to Gdm^+ and Cl^-.

Fig. 5.7 shows the values for π_{XY} where X is the urea nitrogen atom or the urea oxygen atom on the top and bottom rows respectively. The Y sites refer to different backbone and sidechain sites on each of the three peptides. Panel a in Fig. 5.7 shows evidence for accumulation (π_{XY} ¿ 1) of the nitrogen atoms of urea molecules around each carbonyl oxygen atom of the poly-glycine backbone. The magnitudes of π_{XY} are similar around the different sites along the chain. There is a depletion of the nitrogen atoms of urea molecules around

the amide nitrogen atoms of the backbone. The values of π_{XY} are approximately unity for the oxygen atoms of urea around the carbonyl oxygen and amide nitrogen atoms of the backbone. This implies a lack of accumulation or depletion of urea oxygen sites around the polyglycine backbone - see panel d in Fig. 5.7.

Panels b and e of Fig. 5.7 show the π_{XY} values obtained for the relative occupancies of urea oxygen (panel b) and urea nitrogen (panel e) atoms around backbone and sidechain sites of the CAP peptide. These plots show increased variation in the values of π_{XY} around backbone sites when compared to what we calculate around similar sites for polyglycine. Secondly, the accumulation of urea nitrogen atoms around specific sidechain sites is equivalent to or higher than the accumulation of urea nitrogen atoms around backbone oxygen atoms. These sidechain sites include the primary amide oxygen atoms of Gln and Asn, atoms within the aromatic rings of Phe and Tyr, and atoms of imidazole rings of His. Similar trends are observed for the relative occupancies of urea nitrogen atoms around the backbone and sidechain sites of the OSP peptide. Here, there is accumulation around the carbon atoms of aliphatic sidechains and depletion of the urea nitrogen atoms around the positively charged amines of Lys sidechains. Urea oxygen atoms accumulate around the primary amide nitrogen atoms of Gln and Asn. They also accumulate around the sidechain atoms of Ser and the sites of on Arg and Lys sidechains that carry partial positive charges.

The results shown in Fig. 5.7 can be compared quantitatively with the values for local solute partition coefficients designated as K_P that were recently reported by Diehl *et al.* [138]. Salient agreements are as follows: On average, we obtain π_{XY} values of 1.29, 1.20, 1.1, and 1.04 for the urea nitrogen atoms (X=N) around the backbone oxygen atoms, aromatic carbon atoms, aliphatic carbon atoms, and the hydroxyl oxygen atoms, respectively. These values compare favorably to the corresponding K_P values of Diehl *et al.*, which are 1.28 $\pm$

0.02, 1.28 ± 0.02, 1.03 ± 0.02, and 1.08 ± 0.02 for the interactions of urea with amide oxygen, aromatic carbon, aliphatic carbon, and hydroxyl oxygen atoms, respectively. The central discrepancy between our π_{XY} values and the K_P values arise for the interaction of urea with amide nitrogen atoms. We obtain an average value of 0.9 for π_{XY} where X=O for the interaction of urea oxygen atoms around the backbone amide nitrogen of G_{15} whereas Diehl et al. report a K_P value of 1.10 ± 0.07 for the interaction of urea with backbone amide nitrogen atoms. The disagreement is greater when we consider the average π_{XY} value of 0.64 for the interaction of urea oxygen atoms around the backbone amide nitrogen atoms of CAP and OSP, respectively. This discrepancy originates mainly from the effects of chain connectivity and occlusion of the backbone amide nitrogen by the sidechains in CAP and OSP, and both these features are absent in the model compounds used to arrive at partition coefficients.

5.4 Discussion

We begin the discussion with a summary of the results: Polypeptide backbones form compact globules in water. The preference for compact globular conformations persists in high concentrations of denaturants although modest expansion derives from the sampling a more swollen globule with occasional but transient globule-coil-globule transitions. Therefore, the observed expansion of generic protein sequences in highly denaturing environments cannot simply be attributed to preferential interactions of denaturants with backbone moieties [53]. We uncover a two-stage mechanism to explain the effect of sidechains on protein denaturation. In water, in the absence of denaturants, favourable sidechain-solvent interactions induce a dilution in the effective concentration of polypeptide amides. Further accumulation of denaturant molecules around backbone and sidechain sites, in accord with the solute partitioning model and observations from detailed as well as coarse grained molecular dynamics simulations, leads to expansion that results in conformational properties that become congruent with those of canonical random coils [87, 88, 124, 129, 427, 430, 476, 658].

Our results highlight the need to consider the thermodynamic impact of the three-way competition among amide-amide, amide-water, and amide-denaturant interactions. In the absence of sidechains, the effective amide-amide interactions are stronger than the totality of the effects of amide-water and amide-denaturant interactions. Consequently, while the values associated with the interaction parameter π_{XY} are in accord with the partition coefficients summarized by Diehl *et al.* for urea, these values alone do not help in quantifying the extent of chain expansion that is realized for a protein sequence [138]. This is because the effects of chain connectivity on the effective amide-amide interactions cannot be incorporated into estimates based on model compounds. Our results suggest that the energy scales for effective amide-amide interactions are weakened by sidechains, which act as a local solvent matrix

for backbone amides. This, sidechain priming effect, when combined with the additive contributions from preferential interactions of denaturant molecules with specific protein sites will give rise to chain expansion that is consistent with the statistical properties of canonical random coils. Our work highlights the importance of quantifying the effective concentration of backbone amides. This quantity, unlike solvent accessible surface areas, might be a useful descriptor of the effects of conformational properties because it can be converted into an estimate of the effective amide-amide interactions given knowledge of the energetics of amide-water and amide-denaturant interactions.

5.4.1 The Role of Glycine Patterning and Context

A prediction that emerges from this interpretation is that while long contiguous stretches of poly-glycine would be expected to compact via backbone-mediated amide-amide interactions, glycine-rich regions interspersed with non glycine residues would lack the ability to form dense glycine-rich globules due to the presence of sidechains, effectively blocking efficient backbone-backbone interactions and engendering significant chain expansion. Importantly, when unable to drive compaction via amide-amide interaction, glycine might be expected to facilitate chain expansion through enhanced flexibility and the absence of a sidechain to mediate inter-residue interactions. While sidechain-backbone interactions are certainly expected for some residues, the geometries associated those interactions are significantly more restrictive than a sidechain-sidechain interaction. Taken together, we expect glycine's impact on chain behaviour to have a highly context dependent role. In folded proteins it is likely to be found at helix caps, in turns, in poly-proline II helices, or in flexible loops. In disordered proteins it may be found in flexible, expanded sequence or in more compact

polar rich tracts where it engenders weak amide-amide interactions while facilitating chain flexibility.

Recent work by Gates et al. provides an interesting test case for this hypothesis [193]. In this work, the authors report on the biophysical characterization of a curious 81 residue glycine-rich region of the snow-flea antifreeze protein (sfAFP). Despite lacking a hydrophobic core, this sequence folds into a stable polyproline-II rich fold, held in place by two disulphide bonds (see chapter 2 for a rendering of the folded structure). Under reducing conditions sfAfp is unable to fold, and despite a glycine content of ∼45% forms a an expanded ensemble with an average R_G of 23.1 ± 0.1 Å. This result is broadly consistent with the hypothesis that long (¿ 10 - 15) contiguous runs of glycine are required for dense glycine-glycine interactions to drive compaction. We sought to determine if the results described by Gates et al. were still consistent with a model where polyG undergoes collapse. We used a preliminary version of the General Chemical Forcefield with PIMMS (introduced and described in chapter 14 to generate ensembles of the 81 residue sfAFP (amino acid sequence CKGADGAHGVNGCPGTAGAAGSVGGPGCDGGHGGNG-GNGNPGCAGGVGGAGGASGGTGVGGRGGKGGSGTPKGADGAPGAP) as well run equivalent simulations for G_{81}. As shown in 5.9, while polyG collapse into a dense globule sfAFP forms a coil-like ensemble with globale dimensions entirely consistent with available SAXS data, highlighting the strong sequence-context dependence associated with glycine.

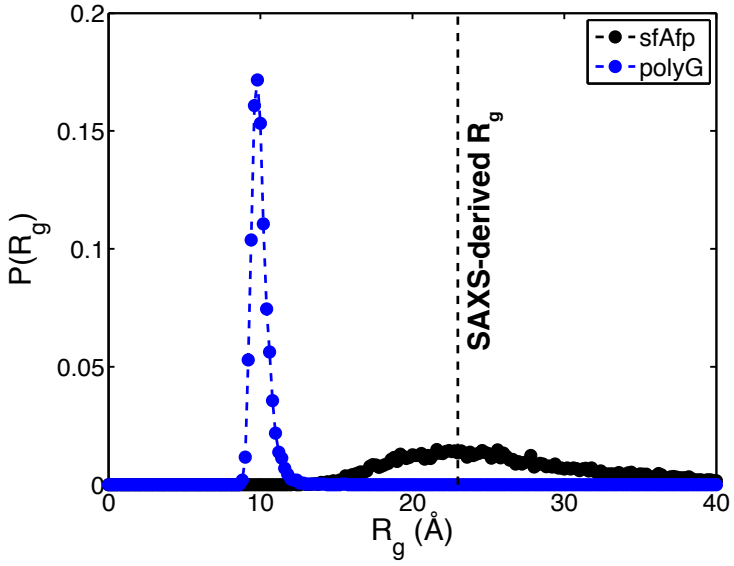

Figure 5.9: R_G distributions for ensembles of polyG and sfAfp generated using PIMMS. Despite the high glycine content sfAfp exists in a coil-like ensemble, while polyglycine forms a compact globule. Both results are generated with the same forcefield, and are consistent with experimental characterizations of polyglycine (this work) and sfAfp [193].

5.4.2 Impact of Forcefields for Denaturant Molecules

Tran *et al.* used parameters from the OPLS-AA forcefield to model the effects of high concentrations of urea on the conformational properties of polyglycine [138]. The combination of the KBFF forcefield for urea and TIP3P for water molecules reproduces the near ideality of urea-water mixtures across the entire solubility range of urea [499, 634–636]. In contrast, the combination of OPSL-AA and TIP3P shows considerable non-ideal clustering of urea molecules [499, 634]. This points to inaccuracies in the balance of solute-solute, solute-solvent, and solvent-solvent interactions with the OPLS-AA forcefield. These inaccuracies engender stronger clustering of urea molecules around polypeptide amides, which

leads to significant chain expansion that is inconsistent with our simulation results based on the KBFF forcefield and our FCS data.

5.4.3 Connections to Interpretations from the Transfer Model

Data regarding the denaturant dependence of solubility of backbone and sidechain analogs have been used to develop mechanistic inferences regarding protein denaturation [18, 19]. According to a specific version of the transfer model, preferential interactions with backbone amides provide the main driving force for denaturation in urea. In this interpretation, the picture that emerges is one of a backbone centric view for protein denaturation with sidechains playing a passive role [53]. Our results indicate that pure polypeptide backbone constructs, devoid of sidechains, undergo modest expansion. Therefore, preferential interactions of urea with the backbone cannot explain the extent of denaturation measured for generic protein sequences. Further, we demonstrate the priming of the backbone in the absence of denaturants and we implicate this intrinsic expansion in water as a contributor to protein denaturation. The results in Fig. 5.7 demonstrate that the primed backbone units interact differently with urea when compared to the backbone units devoid of sidechains. Overall, our findings are consistent with those reported by Moeser and Horinek [397]. They used molecular dynamics simulations to assess the accuracy of the backbone centric version of the transfer model. Moeser and Horinek found significantly improved correlation between the transfer free energy and change in solvent accessible surface area upon unfolding when they use a 'universal backbone' construct. This construct accounts for synergy between the backbone and sidechain moieties in the form of a "compensating error" in the transfer free energies of sidechain groups. In effect, Moeser and Horinek demonstrate that one can

construct an additive transfer model if one were to account for synergistic rather than independent contributions of backbone and sidechain moieties to interactions with urea. These findings are conceptually congruent with our results, although we take a different route toward uncovering a mechanistic interpretation of the origins of preferential interactions.

Recently Wei *et al.* reported simulation results, obtained using AMBER99 forcefield for peptides, the SPC/E water model, and the OPLS-AA forcefield for urea [637]. These results point to sidechain-specificity in the sequential destabilization of backbone hydrogen bonds of beta hairpins. As noted above, the OPLS-AA forcefield shows considerable non-idealities in terms of anomalous clustering of urea molecules that engender spuriously strong interactions of urea with peptide amides as well. Therefore, we see the results of Wei *et al.* as being in qualitative agreement with the with the two-stage mechanism that we propose based on our results.

5.4.4 Reconciling Competing Models for Denaturant-Protein Interaction

Over the last fifty years multiple models have been proposed to describe the mechanism through which denaturants interact with polypeptide. In the earlier years three putative models were proposed; (1) Denaturants disrupt the inherent structure of water on a macroscopic level, leading to protein unfolding (2) Denaturants interact with sidechain groups, driving chain expansion (3) Denaturants interact with the backbone, driving chain expansion. The water-structure argument fell by the wayside as increasingly detailed experiments and simulations found no evidence for such a model [299, 481, 553]. However, the backbone vs. sidechain models have both received continued attention until recently.

While various other reports has shown that denaturants interact with both backbones and sidechains, we feel that our work provides a satisfying explanation as to *why* both of these interactions are necessary for denaturation [87, 88, 241, 397]. The urea-sidechain interactions are a necessary component to allow urea-backbone interactions to occur. The urea-backbone interactions are critical for denaturation, and our results agree with even the most staunch backbone-only arguments, yet they are only possible in the presence of sidechains. Urea sidechain interactions, in turn, further drive backbone accessibility, in effect giving rise to local, psuedo-cooperative unfolding. GdmCl strongly interacts with sidechains, expanding the chain and allowing backbone to engage in backbone-water interactions because backbone-backbone interactions are no longer a viable option. This mechanism would be expected to show a strong two-stage kinetic behaviour, with sidechain-Gdm interaction leading to protein expansion followed by backbone solvation and denaturation. In a perfectly agreement with our results, this is the precise mechanism reported by Jha and Marqusee [265].

5.4.5 Reconciling Our Observations With the SAXS Data of Kohn *et al.*

Our results for the conformational properties of the backbones of CAP and OSP in 8 m urea and 8 m GdmCl are congruent with the highly denatured state behaving like the FRC limit, rather than then EV limit. At first glance, this seems to be at odds with the scaling of R_G with N that is derived from SAXS and single molecule spectroscopy. There are four reasons for the discrepancy: (i) We compare the statistical properties of polypeptide backbones to those observed in reference ensembles for sequences with and without sidechains. Therefore, part of the disagreement originates in the fact that SAXS data for R_G include contributions from the scattering cross-sections of sidechain and backbone atoms. (ii) The finite size of

CAP and OSP - they are 15-residue fragments as opposed to being *bona fide* full-length sequences - is another reason for the discrepancy between simulation results and the inferences of Kohn *et al.*. For longer chains, the amino acid compositions within polymeric segments along the sequence will, on average, be in accord with the biases seen in globular proteins. Increased sidechain priming and the increased number of sites for denaturant accumulation should yield dimensions that match those observed in experiment. (iii) Meng *et al.* recently showed that an exponent of $\sim$0.59 in high concentrations of urea is compatible with quantifiable deviations from the conformational properties in the EV limit [377]. Indeed, further work on NTL9 demonstrates that scaling at 0.59 is compatible with a range of strong, anisotropic intramolecular interactions (see chapter 7). Although mean R_G values for highly denatured proteins scale $\propto N^{0.59}$, the actual R_G values are considerably smaller than those expected from the EV limit, and this discrepancy increases with increasing chain length. Therefore, residual intra-chain attractions do prevail even in apparent good solvents. Meng *et al.* attribute these to low-likelihood non-native clusters of hydrophobic residues. Consequently, the degree of expansion beyond the FRC limit is actually rather modest for proteins in aqueous solutions with high concentrations of urea or GdmCl. (iv) Finally, our results suggest a higher degree of expansion for the backbone of OSP over that of CAP in 8 *m* GdmCl. This points to possible weaknesses of the KBFF forcefield in capturing cation-pi interactions that are expected to be important for denaturation in high concentrations of GdmCl.

5.4.6 Unfolded States Under Folding Conditions

Our results suggest that sidechain prime the backbone for expansion by diluting the effective concentration of amides even in the absence of denaturant molecules. This observation leads

us to propose a two-stage mechanism for protein denaturation that highlights the importance of sidechains, not just in their interactions with denaturants, but also as determinants of the conformational properties of unfolded states in the absence of denaturants. It is noteworthy that early work based on nuclear magnetic resonance spectroscopy and stopped flow kinetics yielded evidence demonstrating that the unfolded state under folding conditions is clearly distinct from the ensembles sampled by generic proteins in high concentrations of denaturants [14,92,262,398,669]. Our findings, taken together with results from early studies, raise the question of the effective exponent ν_{eff} that best describes the scaling with chain length of the dimensions of unfolded ensembles in the absence of denaturants. The transfer model implicitly stipulates that $\nu_{\text{eff}} \approx 0.59$, especially for proteins that show apparent two-state behaviour [476,592]. A second alternative is that $\nu_{\text{eff}} \approx 0.33$ implying that unfolded ensembles under folding conditions follow the properties of polypeptide backbones in water. Neither of these alternatives are supported by our results.

A recent collection of published results, and our own work on NTL9 are particularly pertinent with respect to this question [20, 59, 234, 672, 673]. In these reports, a variety of approaches was used to estimate the values of ν_{eff} for the unfolded ensembles under folding conditions (or close to) for several different proteins. Hofmann *et al.* used single molecule FRET (smFRET) on a range of different folded and disordered proteins [234]. Aznauryan *et al.* combined smFRET with NMR and SAXS results to examine the unfolded state of ubiquitin [20]. Borgia *et al.* combined smFRET with SAXS, two focus FCS, and dynamic light scattering to examine a destabilized mutant of the spectrin R17 domain and the intrinsically disordered protein ACTDR [59]. In our work in chapter 7 we combined time resolved SAXS, time resolved FRET, and all atom simulations to describe the transient unfolded population of NTL9 under folding conditions. Single molecule spectroscopy affords the resolution to separate folded and unfolded populations under folding conditions. This allows one to follow

the evolution of conformational properties of unfolded states as a function of denaturant concentration. Alternatively, time-resolved methods allow a similar distinction to be made for proteins where the unfolded state is a short-lived and transient species. These are fundamentally distinct methods in terms of how they obtain the unfolded state under folding conditions.

The general consensus from all these studies is that the collapse transition is broadly continuous, although the unfolded state under folding conditions is still relatively expanded compared to the folded state. This observation is apparently contradicted by inferences from SAXS measurements, although several of the papers published here, suggest explanations for this long-standing discrepancy (also see chapter 8 for additional discussion). Regardless of the exact scaling exponent, the overwhelming evidence is that the unfolded ensemble under folding conditions is distinct from the denatured state ensemble sampled under highly denaturing conditions - a finding that agrees well with earlier studies [14, 92, 262, 398, 669]. Ensemble measurements of several marginally stable proteins and high-throughput simulations based on distributed computing have yielded similar conclusions regarding the non-equivalence of unfolded states under folding conditions versus those sampled in highly denaturing or unfolding environments.

Of direct interest and relevance are the estimates for ν_{eff} for the unfolded state under folding conditions. Hofmann *et al.* suggest that ν_{eff} ranges from 0.4 to 0.51 depending on the overall hydrophobicity and charge content of the underlying sequence [234]. Aznauryan *et al.* estimate $\nu_{\text{eff}} \approx 0.5$, and a similar value is obtained by Borgia *at al.* [20]. In our work we also estimate $\nu_{\text{eff}} \approx 0.5$, but note that, broadly speaking values from 0.5 to 0.55 are consistent with the available data, given that even for ensembles that display $\nu_{\text{eff}} \approx 0.55$

strong deviations from the EV ensemble are observed, including well defined local and long-range interactions and well defined (albeit transient) formation of secondary structure.

For a two-stage mechanism of unfolding, the value for ν_{eff} prescribes the degree of intrinsic expansion and hence the extent of dilution that needs to be achieved in order to realize an exponent of ∼0.59 in denaturing environments. If we set $\nu = 3/5$ as the target for the scaling exponent in highly denaturing environments, then the extent of dilution needed to be achieved will scale as $N^{1.8-\nu_{eff}}$ with chain length, providing the degree of intrinsic expansion for unfolded states under folding conditions is quantified using ν_{eff} [504]. The intrinsic expansion of backbones in solutions with high concentrations of denaturants is rather modest. Accordingly, the values for ν_{eff}, as dictated by amino acid composition, would have to be in the range of 0.5 if generic denatured state ensembles are to have dimensions that are congruent with a scaling exponent of ≈ 0.59.

5.4.7 Most Proteins Show Similar Amino Acid Compositional Biases

In light of NMR, SAXS and single molecule data for the scaling exponent that characterizes the dimensions of highly denatured proteins, we propose that proteins that have been subjected to scaling analysis in high concentrations of denaturants have similar amino acid compositional biases. We used a simplified alphabet and divided amino acids into disorder promoting (Ala, Arg, Asp, Gln, Glu, Gly, His, Lys, Ser, Pro, Thr) versus order promoting (Asn, Cys, Ile, Leu, Met, Phe, Trp, Tyr,Val) sets [86, 153]. This partitioning is reminiscent of the 'HP-code' of Chan & Dill [93]. We find that the ratio of disorder to order promoting residues is 64:36 for proteins in the dataset of Kohn *et al.* [297]. This ratio is 62:38 for

sequences of single domains drawn from the PSBSelect25 database of non-redundant protein sequences [205]. The implication is that the compositions of generic protein sequences support the tenets of the proposed two-stage mechanism. Accordingly, there will always be a sufficient fraction of sidechains to prime the backbone for expansion of unfolded states in water thus giving rise to values of ν_{eff} that are around 0.5. The generic sidechain compositional biases within most protein sequences therefore encodes the possibility of counterbalancing of intra-chain and chain-solvent interactions for unfolded states in the absence of denaturants. This should give rise to statistical properties for unfolded states under folding conditions that are congruent with those of polymers in Θ solvents [180, 181].

5.4.8 Foldable Proteins Sequences Select for Metastability

A temping (albeit entirely untested) hypothesis is this balance of disorder and order promoting residues provides a means to encode metastability into folded proteins. Proteins that are enriched in order promoting residues may be at greater risk of aggregation during folding, misfolding into inescapable folding intermediates, or causing insurmountable thermodynamic challenges for the cellular proteostatic machinery. Conversely, proteins enriched in disorder promoting residues may be unable to fold, as demonstrated by intrinsically disordered proteins. The ratio of ∼60:40 disorder:order may encode Θ-like behaviour in the unfolded ensemble, facilitating the acquisition of the native state by allowing the chain to efficient search through conformation space through a loose nucleation-collision style folding mechanism (see chapter 1 for further detail on folding mechanism). While we include charged residues in the 'disorder' promoting category here, there are many examples of highly-charged folded proteins that show extreme stability due to the topological constraint imparted by the charged and hydrophobic residues. Consequently, the relationship between composition

and metastability is likely to be more complex than simply relative fractional content of amino acids. Never-the-less, the idea that an additional constrain on protein evolution that is somewhat orthogonal to its ability to form a stably folded protein (and would be largely invisible to conventional *in vitro* activity assays) is an interesting prospect, and warrants further study.

Chapter 6

Sequence Determinants of the Conformational Properties of an IDP Prior to and Upon Multisite Phosphorylation

The following section is taken from the paper **Sequence Determinants of the Conformational Properties of an Intrinsically Disordered Protein Prior to and upon Multisite Phosphorylation** by E.W. Martin, A.S. Holehouse, C.R. Grace, A. Hughs, R.V. Pappu, and T. Mittag. This was published in the *Journal of the American Chemical Society*, Vol. 138, pages 15323 - 15335, in November 2016. The text has been expanded to include additional detail. All experimental work was performed by E.W.M, C.R.G, and A. H. (*not A.S.H.*). A.S.H. performed all simulations, generated sequences designs, and developed simulations-based analysis techniques.

6.1 Background

The downstream responses of cells to different cues are often controlled by signals that are initiated by post-translational modifications. These include multisite Ser / Thr / Tyr phosphorylation within intrinsically disordered regions (IDRs) of specific proteins [252]. Multisite phosphorylation is dynamic and provides a putative mechanism for rapid signal integration [110, 278, 560]. Many nonlinear downstream responses such as transcriptional regulation, cell cycle control, and cell proliferation are coordinated by multisite phosphorylation of IDRs [58, 122, 137, 213, 238, 298, 331, 357, 449, 566, 611]. Archetypal IDRs that undergo multisite phosphorylation include the C-terminal domain of RNA polymerase II, the C-terminal tail of the epidermal growth factor receptor, and sidearms of intermediate filaments in neurons [439, 449, 460, 670].

Sites of phosphorylation are often located within short linear motifs (SLiMs) [609]. These motifs are the substrates for kinases and phosphatases that catalyze site-specific phosphorylation and dephosphorylation, respectively [608]. The accessibilities of substrate motifs to kinases (writers), downstream binding partners (readers), and phosphatases (erasers) are governed by sequence-encoded local and global conformational properties of IDRs prior to and following multisite phosphorylation. The overall fraction of charged residues (FCR) increases upon multisite phosphorylation. If the net charge per residue (NCPR) prior to phosphorylation is close to zero, then multisite phosphorylation will induce a polyampholyte to polyelectrolyte transition. Conversely, if the NCPR is larger than zero, then multisite phosphorylation will induce a transition from a polyelectrolyte to a polyampholyte, and the sequence patterning of oppositely charged residues is expected to become an important determinant of conformational properties [126]. Therefore, multisite phosphorylation has the potential to induce significant changes to the conformational properties of IDRs.

The sequence-encoded balance between protein-solvent and intra-protein interactions determines the conformational properties of IDRs [127, 359, 364, 405]. Recent studies have combined results from all atom simulations and *in vitro* experiments to uncover how sequence encodes the balance between intra-chain and chain-solvent interactions. These findings support a grouping of IDRs into distinct conformational classes based on their amino acid compositions and the sequence patterning of oppositely charged residues [126]. This classification applies to sequences of IDRs that are deficient in hydrophobic and proline residues and are enriched in polar and / or charged residues. It is noteworthy that SLiMs encompassing phosphosites often include proline residues, and that proline-directed kinases regulate a larger number of proteins than non-proline directed kinases [345, 598]. Consequently, many IDRs that undergo multisite phosphorylation will include a moderately high fraction of proline residues. Proline is unique in being an imino acid, giving it distinct structural properties when compared to the amino acids. It disrupts the propagation of regular secondary structural elements, helps nucleate α-helices, promotes turn formation, engenders local stiffening of the backbone, encodes a distinct preference for locally expanded polyproline II conformations when the peptide bond is in the trans configuration, engenders a preference for positive backbone -angles for residues directly N-terminal to it, and can promote global compaction via *trans* to *cis* isomerization [17, 115, 478, 497, 520].

Here, we go beyond previous descriptions of composition-to-conformation relationships for IDRs to investigate the interplay amongst proline, charged, and post-translationally modifiable Ser / Thr residues as determinants of conformational properties of an archetypal IDR prior to and following multisite phosphorylation [126, 127]. The IDR of interest is derived from the protein Ash1, a transcription factor that regulates mating type switching in *S. cerevisiae* [111]. In Ash1 a C-terminal zinc finger domain binds DNA while the remainder of the sequence is predicted to be disordered. It contains 23 Ser / Thr phosphosites that are part

of distinct proline-containing SLiMs. These Ser / Thr residues are phosphorylated by the cyclin-dependent kinases Cln1,2/Cdc28 [341]. Ash1$^{420\text{-}500}$, which is the object of our study, is an 81-residue section of the IDR. It encompasses ten Ser / Thr residues within phosphosites, sixteen arginine and lysine residues, and one aspartate residue (fig. 6.1). The FCR and NCPR of Ash1$^{420\text{-}500}$ prior to phosphorylation are 0.2 and +0.18, respectively. Stoichiometric multisite phosphorylation should change the NCPR to +0.06 while increasing the FCR to 0.35. This change converts the sequence of Ash1$^{420\text{-}500}$ from a weak polyelectrolyte to a well-mixed strong polyampholyte. Accordingly, heuristics that do not account for the contribution of proline residues suggest that multisite phosphorylation would convert Ash1$^{420\text{-}500}$ from a globule to a swollen, well-solvated coil. If this is valid, then there should be a substantial increase in the radius of gyration (R_G) upon multisite phosphorylation [126, 359].

We quantified the conformational properties of Ash1$^{420\text{-}500}$ using Small Angle X-ray Scattering (SAXS), Nuclear Magnetic Resonance (NMR) spectroscopy, and all atom simulations. These studies reveal that unphosphorylated Ash1$^{420\text{-}500}$ adopts expanded coil-like conformations in aqueous solutions. These conformational preferences persist upon stoichiometric as well as sub-stoichiometric multisite phosphorylation. We identified sequence features within Ash1$^{420\text{-}500}$ that determine its intrinsic conformational properties. Specifically, we show that the apparent insensitivity of global dimensions upon phosphorylation derives from compensatory conformational changes along the sequence of multiply phosphorylated Ash1$^{420\text{-}500}$.

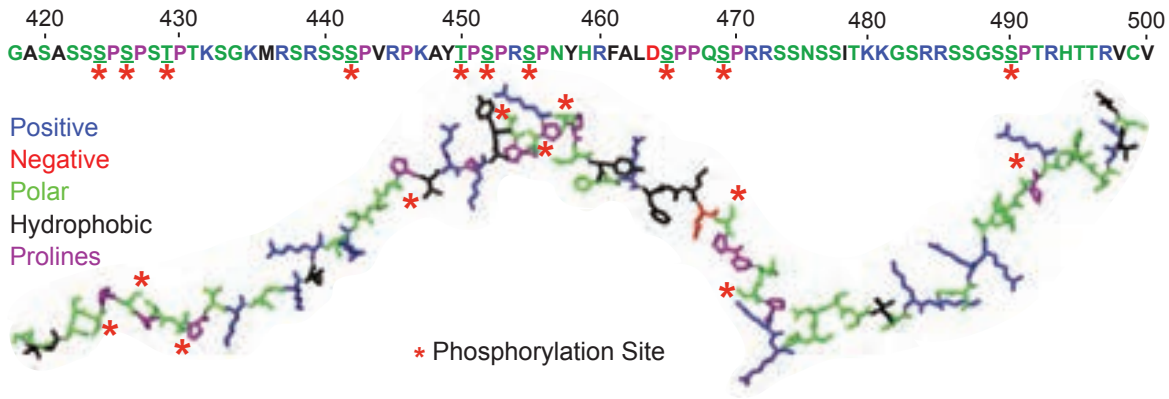

Figure 6.1: The sequence of Ash1$^{420\text{-}500}$ is shown at the top and the color-coding of residues is described below. The residues are also shown in a stick representation for a generic conformation of Ash1$^{420\text{-}500}$. Phosphorylation sites are underlined in the primary sequence and marked by red asterisks in the conformation. Positively and negatively charged residues, small polar residues, hydrophobic and proline residues are colored blue, red, green, black and purple, respectively. The Ash1$^{420\text{-}500}$ construct has two exogenous N-terminal residues, GA, and these remain after cleavage of the affinity tag.

6.2 Methods

6.2.1 Protein Expression and Purification

His-tagged Ash1$^{420\text{-}500}$ and 5pAsh1$^{420\text{-}500}$ was expressed in an *E. coli* BL21 GOLD (DE3) strain (Agilent) in LB or M9 media for isotope-labelled samples. Expression was induced at $OD_{600} = 0.8$ with 0.6 mM IPTG and cells were cultured at 20°C for an additional 18 hours. His_6-Ash1$^{420\text{-}500}$ was purified from inclusion bodies. The polyhistidine tag was cleaved with a TEV protease, which left the protein with two additional N-terminal residues, i.e., Gly-Ala. We refer to this protein construct as Ash1$^{420\text{-}500}$. Final protein samples were generated

by size exclusion chromatography on a Superdex 75 column (GE Life Sciences) into the desired buffer. Purified proteins were concentrated using Millipore centrifugal concentrators with 3000 Da cutoff. Their purity, integrity and identity were analyzed by SDS PAGE gel, MALDI-TOF and LC-MS/MS. The concentration was assessed via absorbance at 280 nm ($\epsilon = 2980\,\text{M}^{-1}\,\text{cm}^{-1}$).

6.2.2 Protein Phosphorylation

Phosphorylated samples were prepared by treatment of Ash1$^{420\text{-}500}$ with Cyclin A/Cdk2 (prepared according to Huang et. al55) at a kinase/Ash1$^{420\text{-}500}$ ratio of 1:100 in the presence of 50 fold excess of ATP and 2.5 mM $MgCl_2$ overnight at 30 °C. Substochiometric ratios of ATP to Ash1 of 12.5 and 5 were used to generate Ash1 populations with distributions centered around 5 and 2 phosphorylated sites, respectively. The yield of the phosphorylation reaction was determined by ESI-TOF mass spectrometry.

6.2.3 SAXS Sample Preparation and Data Collection

Samples of Ash1$^{420\text{-}500}$ were prepared in a buffer containing 50 mM Tris pH 7.5, 10 mM DTT and 2 mM TCEP. High concentrations of Tris and DTT were used to scavenge radicals and prevent radiation damage. The addition of TCEP served to stabilize the buffer reduction potential over the course of shipping and waiting for measurement. Purified protein samples were concentrated to approximately 2 mM and were then diluted into buffers to achieve the desired NaCl concentrations.

Solution SAXS data were collected at both the 12-ID-B beamline at the Argonne National Laboratory Advanced Photon Source and through the mail-in program at the SIBYLS beamline at the Lawrence Berkeley National Laboratory Advanced Light Source. SAXS data were acquired manually at APS, where protein samples were loaded, then gently refreshed with a syringe pump to prevent x-ray damage. A Pilatus 2M detector provided q-range coverage from 0.015 $Å^{-1}$ to 1.0 $Å^{-1}$. Wide-angle x-ray scattering data were acquired with a Pilatus 300k detector and had a q range of 0.93 - 2.9 $Å^{-1}$. Calibration of the q-range calibration was performed with a silver behenate sample. Protein samples were freshly prepared using size exclusion chromatography (GE Life Sciences, Superdex 75 10/300 GL) in a buffer containing 50 mM Tris pH 7.5, 150 mM NaCl, 5 mM DTT, and 2 mM TCEP. Elution fractions were loaded without further manipulation. Buffer collected 1 column volume after protein elution from the column was used to record buffer data before and after each protein sample. Twenty sequential images were collected with 1 sec exposure time per image with each detector. Data were inspected for anomalous exposures and mean buffer data were subtracted from sample data using the WAXS water peak at q $\sim$ 1.9 $Å^{-1}$ as a subtraction control.

Samples were sent to SIBYLS in 96 well plates (VWR). A pipetting robot automatically exchanged samples. SAXS data were measured for samples at protein concentrations of 450, 225, and 112 µM for each NaCl concentration. Matched buffers were collected from centrifugal concentrator filtrate and were included in wells before and after each dilution series. Data were collected in a q-range of 0.012 - 0.324 $Å^{-1}$ using 0.5, 1, 2 and 5-second exposures. Buffer-subtracted data from each exposure time were manually assayed for high noise and radiation damage. Data were then merged into a single data set using the program PRIMUS [300].

6.2.4 SAXS Data Analysis

Basic analysis including raw data plotting, Kratky transformations to determine flexibility and Guinier transformations to estimate R_G were performed with the program ScÅtter or in-house written MATLAB scripts. Care was taken to limit the Guinier region to very low q values suitable to a disordered protein system. The form factors of IDP ensembles will span the range between rods and spheres implying that the appropriate q-range maximum for Guinier analysis should lie between $q \times R_G = 0.7 - 1.4$ [183, 568]. The best region was chosen by minimizing deviations in the calculated R_G due to either the removal of points near the beam stop or inclusion of higher q points. Ensemble modelling of SAXS data was done using the Ensemble Optimization Method (EOM 2.0) in the ATSAS software package in which a genetic algorithm is used to select an ensemble of conformations from a randomly generated pool [39]. Pools of Ash1 conformations used were alpha carbon traces created by EOM. Ensemble R_G distributions obtained for all salt concentrations were fit to a function describing the R_G distribution of a non-intersecting chain in three dimensions [328, 389].

6.2.5 NMR Data Collection

NMR data were acquired on Bruker Avance 600 and 800 MHz spectrometers equipped with TCI triple-resonance cryogenic probes and pulsed-field gradient units. All samples were prepared in an NMR buffer consisting of phosphate-buffered saline (137 mM NaCl, 2.7 mM KCl, 10 mM Na_2HPO_2, 1.8 mM KH_2PO_4), 10 mM DTT pH 6.95 and 10% D_2O at 5°C. For assignment, approximately 0.7 mM ^{15}N, ^{13}C $Ash1^{420\text{-}500}$ sample was used to acquire standard triple-resonance backbone assignment experiments and carbon-detect triple resonance experiments. Standard assignment experiments were based on sensitivity enhanced ^{1}H-^{15}N

HSQC (8 scans, 2048 × 320 complex data points, with 12 ppm and 25 ppm as ^{1}H and ^{15}N sweep widths). Carbon detect experiments were based on (HA Start) CON-IPAP (16 scans, 1024 × 512 complex data points, with 18 ppm and 36 ppm as ^{13}C and ^{15}N sweep widths) [32].

^{15}N NMR relaxation experiments acquired on a Bruker Avance 800 MHz spectrometer at 278 K using standard pulse programs (16 scans, 2048 (^{1}H) × 150 (^{15}N) complex data points). The longitudinal R_1 spin-lattice relaxation rates were measured using relaxation delays of 20, 50, 200, 500, 1000, 1500, 2000 and 3000 ms. Transverse R_2 spin-spin relaxation rates were measured using relaxation delays of 92.5, 185, 277.5, 370, 462.5, 555, 740 and 925 ms. Relaxation rates were determined by integrating peak amplitudes and fitting to a single exponential decay. Error values are determined via 95% confidence intervals calculated using the residuals and Jacobian matrix from the nonlinear fit.

Data were processed using BRUKER Topspin version 3.2, NMRPipe (v.7.9) and analysed using CARA (v.1.8.4) [135, 284]. All spectra were referenced directly using DSS for the ^{1}H dimension, ^{13}C and ^{15}N frequencies were referenced indirectly. Secondary structural propensities were calculated using ^{13}C', $^{13}C\alpha$, and $^{13}C\beta$ chemical shifts and the SSP and ncSPC algorithms [363, 570]. PPII propensities were calculated from the same pool of chemical shifts with the addition of N, and $^{1}H^{N}$ shifts using δ2D [85].

6.2.6 All Atom Monte Carlo Simulations

All simulations were performed using the CAMPARI Monte Carlo modelling suite. The simulations deploy the ABSINTH implicit solvent model and forcefield paradigm [613]. The protein atoms and mobile solution ions are modelled in atomistic detail, while the solvent

is treated using a mean field (implicit) model. Move sets combine pivots, concerted rotations, sidechain rotations, mutual reorientations, translations of the mobile ions, and a series of moves that enable the efficient sampling of the conformational degrees of freedom coupled to proline ring systems [471]. Multiple independent simulations for each construct were run. Accurate modelling of Ash1 conformational equilibria requires the use of suitably optimized parameters for proline residues [471]. Without these parameters, details such as *cis*-to-*trans* isomerization, proper prolyl ring puckering, and the accurate coupling amongst ring puckering, peptide bond isomerization, and backbone phi angles cannot be reproduced. The optimized parameters for proline residues are interoperable with the `abs_3.2_opls.prm` parameter file in CAMPARI. However, since parameters for phosphorylated residues are currently unavailable for this parameter set, we pursued the route of replacing phosphorylated Ser and Thr residues with Glu. This strategy allowed us to investigate the impact of altering the charge distribution upon Ser / Thr phosphorylation, but it should not be viewed as a perfect mimic of Ser / Thr phosphorylation. Simulation analysis was performed using CTraj (see chapter 9), MDTraj and routines built into CAMPARI [374]. Sequence analysis and permutant design was performed using CIDER and localCIDER (see chapter 4).

Aggregate scattering curves were calculated from simulated ensembles using the program CRYSOL in the ATSAS package [569]. Scattering intensities were calculated for individual PDB files and were combined and scaled using MATLAB. In order to generate a random pool of all-atom conformations with statistically validated backbone angles, we used Flexible Meccano and modelled amino acid sidechains using SCCOMP [167,434]. This pool was used to generate a SAXS curve for comparison with experimental data and ABSINTH ensembles.

6.2.7 Proteome-Wide Bioinformatics Screen for Ash1-like Regions

The human proteome was obtained from UniProt, and sequences were annotated for consensus disorder predictions using the D_2P_2 database [426, 600]. Specifically, only regions for which five or more predictors indicated disorder were designated as disordered for further analysis; this is a relatively stringent threshold. Putative phosphosite data for proteome-wide screening were taken from the ProteomeScout database parsed via the ProteomeScoutAPI , although only regions where UniProt annotation also showed multisite phosphorylation were included in this analysis [237]. To limit the analysis to regions equivalent in size to Ash1, we focused on disordered segments equal to or less than 100 residues. While this provides a distinctly conservative estimate of phosphorylation, it ensures that the identified regions come from high-fidelity data.

6.2.8 The Patterning Parameter Ω

The proline/charge patterning parameter Ω reports on the extent of mixing between proline and charged residues with respect to all other residues. If proline and charged residues are well dispersed across an amino acid sequence, that sequence will have a low Ω value. In contrast, if all the proline and charged residues were concentrated in a specific region of the sequence then that sequence would have a high Ω value. Ω is calculated in a manner that is analogous to the calculation of the parameter κ 11 The calculation of κ focuses on the sequence distribution of oppositely charged residues and relies on the fractions and mixing / segregation of positive and negative charges along the sequence. To calculate Ω we use a two-letter alphabet, where each residue is one of either Asp/Glu/Arg/Lys/Pro or one of the

other remaining 15 amino acids.Ω is calculated by first determining the local proline/charge asymmetry with respect to other residues using the patterning asymmetry parameter σ;

$$\sigma = \frac{(f_{+/-/P} - f_{\text{other}})^2}{f_{+/-/P} - f_{\text{other}}} \tag{6.1}$$

Here $f_{+/-/P}$ is the fraction of charged or proline residues within some sequence stretch, while is the fraction of other amino acids (Ala, Cys, Phe, Gly, His, Ile, Leu, Met, Asn, Gln, Ser, Thr, Val, Trp, Tyr) in the same stretch. Ω is calculated over the entire sequence to determine the global local proline/charger asymmetry (σ_G). It is also calculated using a sliding window of 5-6 residues - sequence elements referred to as blobs - to calculate the local patterning asymmetry. The blob size (in residues) is given by the parameter g, and the total number of residues in the sequence is N_{res}. We then calculate the average extent to which the local asymmetry deviates from the global asymmetry, a metric that is quantified by δ;

$$\delta = \frac{\sum_{i=1}^{N_b} (\sigma_i - \sigma_G)^2}{N_b} \tag{6.2}$$

Here, σ_G is the full sequence patterning asymmetry, σ_i is the local patterning asymmetry for blob i, and N_b is the number of blobs in the sequence ($N_b = N_{res}g + 1$). Finally, we introduce a normalization factor, $\delta_{\max}$, which defines the δ value associated with the maximally segregated sequence of the same composition as our sequence of interest. Consequently, Ω is defined as:

$$\Omega = \frac{\delta}{\delta_{\max}} \tag{6.3}$$

As a result, Ω is a normalized parameter that should from 0 to 1. Previous work suggested that the blob length for a peptide lies between 5 and 6 residues. Therefore, for a given sequence, we calculate Ω twice, once with $g = 5$ and once with $g = 6$, and compute the average of the two values. For longer sequences, when $f_{+/-/P} << f_{\text{other}}$, Ω can stray beyond 1.0 [23]. This not a bug, but a function of how Ω is defined. Formally, it is reporting on a normalized description of to what extent local sequence properties match global sequence properties. As a result, for long sequences with a low number of charged/proline residues sequences which are 'well mixed' may deviate more strongly from the sequence average than those that contain a single block of charged/proline residues. Consequently, for sequences where $f_{+/-/P} << f_{\text{other}}$ (or vice versa) Ω may not be a useful parameter. It is worth noting that for these sequences the intrinsic amino acid composition is likely to entirely overwhelm any influence of the residue group group that is in the minority, such that the relative patterning is unlikely to be a useful parameter in these cases.

The calculation of Ω has been implemented in the sequence analysis package localCIDER (http://pappulab.github.io/localCIDER/, see chapter4). This is a high performance framework for the analysis of disordered protein amino acid sequences, and contains a wide array of algorithms for sequence analysis. For more information please see the associated documentation online.

[23] Thanks to Dr. Choi for identifying this edge cases

6.3 Results & Discussion

6.3.1 Ash1 Populates an Expanded Ensemble of Conformations

SAXS data for unphosphorylated Ash1$^{420\text{-}500}$ (referred to hereafter as Ash1) were collected to probe the global conformational preferences of this IDR. Data were recorded at the SIBYLS beamline and the results were independently verified at the Advanced Photon Source at Argonne National Lab, where the data were collected immediately after samples were processed using size-exclusion chromatography (SEC) (see fig. 6.2) for representative data collected in the presence of 150 mM NaCl). The features of the normalized Kratky plots are consistent with an expanded, coil-like ensemble for Ash1 (fig. 6.2). The Guinier regions of the data did not show any indication of aggregation or intermolecular interactions at the concentrations used for SAXS measurements (fig 6.2).

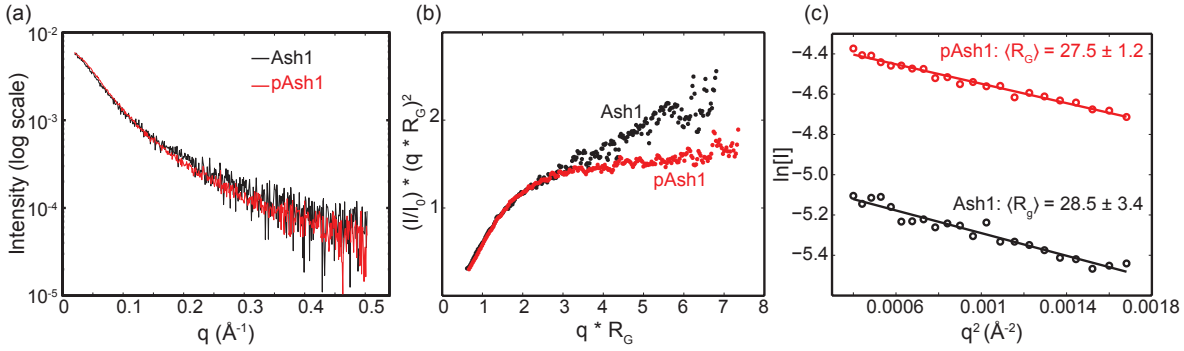

Figure 6.2: Experimental SAXS data indicates disordered nature of Ash1 and pAsh1. (a) Raw SAXS data truncated at $q = 0.5$, (b) dimensionless Kratky plots generated using R_G and I_0 that are calculated from the Guinier analysis and (c) Guinier plots for Ash1 (black) and pAsh1 (red) in aqueous solution and 150 mM NaCl.

A linear fit of the Guinier transformation yields an estimate of the ensemble averaged radius of gyration (R_G). For IDPs, the q-region available for a Guinier analysis is typically smaller than for folded proteins and this was optimized for each sample. We typically analysed q-regions with $q \times R_G < 1$, in agreement with other reports [59,475]. For SAXS data collected immediately after processing by SEC, Guinier analysis yielded an R_G estimate of 28.5 ± 3.4Å for Ash1 in aqueous solutions with 150 mM NaCl. As a reference, a compact globule with the same number of residues would have an R_G of $\sim$ 13 Å. To further analyse the SAXS data, we used the ensemble optimization method (EOM) to generate distributions of radii of gyration that are compatible with the SAXS data for Ash1 [39, 594]. EOM is based on a genetic algorithm whereby a distribution of R_G values is chosen from a randomly generated pool of conformations to ensure that the linear combination of the SAXS profiles of all conformations in the ensemble regenerates the experimental data. Ensembles comprising of 20-30 conformations were typically needed to fit the measured SAXS data (fig 6.3). Additionally, the R_G distribution for Ash1 is shifted to larger sizes with respect to a random starting pool (fig. 6.3b).

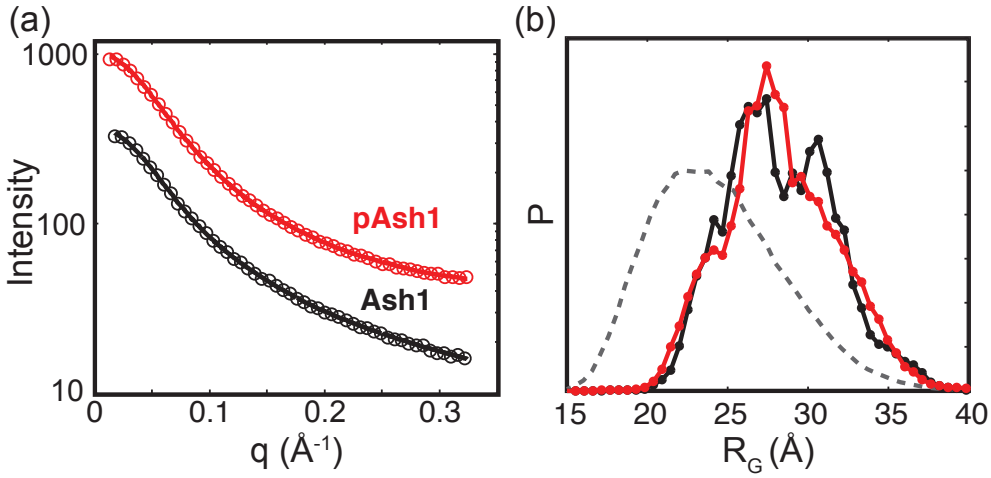

Figure 6.3: (a) Fits of the scattering curves calculated from one representative EOM ensemble to experimental data. Final ensembles are the result of averaging 100 independent iterations. (b) R_G distributions of the random pool (grey, dashed line), and of Ash1 (black line, markers) and pAsh1 (red line, markers) calculated with EOM from SAXS data of samples in aqueous solution containing 150 mM NaCl. The EOM ensembles contain 20-30 conformers, resulting in rough R_G distributions.

6.3.2 Ash1 & pAsh1 Have Similar Global Dimensions

The 10-fold phosphorylated version of Ash1 (referred to hereafter as pAsh1) was generated via overnight incubation of Ash1 with Cyclin A/Cdk2. Analyses of pAsh1 were performed identically to Ash1. In aqueous solutions with 150 mM NaCl, Guinier analysis of SAXS data for pAsh1 yields a mean R_G value of 27.5 ± 1.2 Å. Within error, this value is similar to that of the unphosphorylated Ash1. The EOM analysis yielded R_G distributions for pAsh1 that were similar to those of Ash1 (fig. 6.3b). These results are surprising given the substantial increase in FCR and reduction in NCPR between Ash1 and pAsh1. To assess the robustness of the invariance of global conformational properties to phosphorylation, we generated

an Ash1 mutant with only five intact phosphosites, while the Ser / Thr residues in other phosphosites were mutated to alanine. This construct is referred to as 5pAsh1. Additionally, in an alternative approach, we limited ATP in phosphorylation reactions to generate sub-stoichiometric phosphorylated variants of Ash1 while keeping its sequence intact. In all cases, the average global dimensions and R_G distributions were similar to those of Ash1 and pAsh1 (6.4a), indicating a robustness of the invariance of global dimensions to stoichiometric or sub-stoichiometric phosphorylation [674].

6.3.3 Ash1/pAsh1 Expansions is Insensitive Electrostatic Screening

We reasoned that the net positive charge of Ash1 might engender intra-chain electrostatic repulsions leading to chain expansion. This would be true of archetypal polyelectrolytes. Accordingly, the addition of salt should induce a statistically significant chain compaction through the screening of electrostatic repulsion. To test for this possibility we collected SAXS data for Ash1 over NaCl concentrations ranging from 75 mM to 1500 mM. The overall dimensions, quantified in terms of EOM-generated R_G distributions, were essentially insensitive to changes in salt concentration (fig. 6.4b). The broadening of the R_G distributions at higher salt concentration is likely due to increasingly poor data contrast in SAXS measurements. We find a similar weak sensitivity of R_G distributions to changes in salt concentration for pAsh1 (6.4c). This is true despite a significant increase in the overall charge content (fig. 6.4c). Taken together, these results suggest that in Ash1 and pAsh1 long-range electrostatic interactions are not the main drivers of chain expansion. We sought to further verify this claim by performing a detailed statistical and simulation analysis to determine the extent of change expected if Ash1 were experiencing a polyelectrolyte effect, and compared

the response of Ash1 to salt with that of a similarly length-matched true polyelectrolyte, prothymosin α, as discussed in the following section.

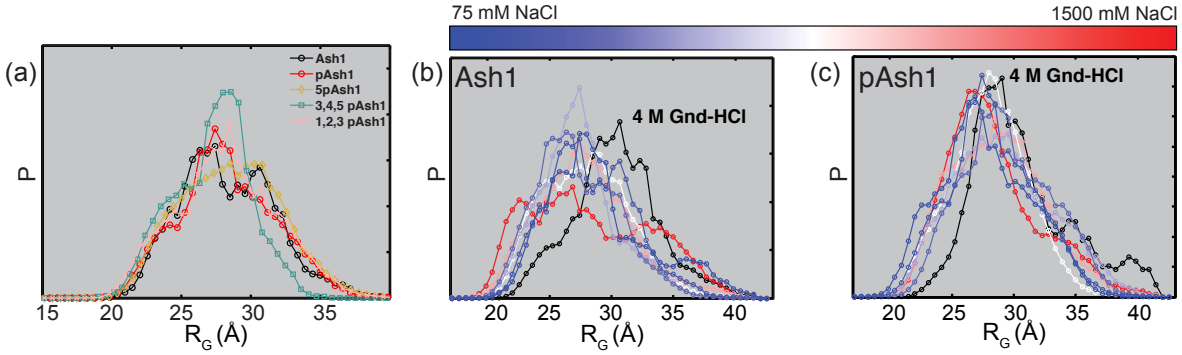

Figure 6.4: The global dimensions of Ash1 are largely insensitive to phosphorylation state and NaCl concentration. (a) Ensemble R_G distributions of Ash1 with different extent of phosphorylation. 5pAsh1 is a mutant with only 5 phosphorylation sites intact, (phosphorylation sites at residues 424, 429, 450, 455 and 469 mutated to Ala), 3,4,5pAsh1 and 1,2,3pAsh1 are generated by kinase treatment with sub-stoichiometric amounts of ATP. Ensemble R_G distributions of (b) Ash1 and (c) pAsh1, respectively, for NaCl concentrations ranging from 75 (blue) to 1500 mM (red) and 4 M Gnd-HCl (black line)

Preferential interactions with denaturants engender further expansion of Ash1 and pAsh1 ensembles, as evidenced by modest increases of R_G values in the presence of 4 M guanidinium hydrochloride (fig. 6.4b and 6.4c). This suggests the presence of weak local structural preferences within the Ash1 ensemble that are lost upon chemical denaturation. Overall, the Ash1 ensembles show a clear preference for coil-like properties and this is true irrespective of the presence or absence of denaturants.

6.3.4 Ash1 Expansion is Not Solely Due to Electrostatic Repulsion

One possible explanation for the large R_G of Ash1 originates from the net positive charge associated with the protein. Given their disordered nature, IDPs frequently have an R_G that is substantially larger than a folded protein of the same number of residues. However, with an R_G of 28.5 Å, the conformational ensemble of Ash1 is approaching that of a peptide-specific perfect self-avoiding random walk (R_G = 33.4 Å), and is ∼2 Å greater than the dimensions predicted for a fully denatured protein of the same length (R_G = 26.3 Å, calculated using $R_0 N_{aa} \nu$, where $[1.927 \times 85^{0.59}]$) as determined based on empirical parameters from Kohn *et al.* [297]. For reference, the discrepancy between the protein-specific perfect random walk and the value predicted by Kohn *et al.* originates from transient long-range interactions in the unfolded state that alter the R_0 prefactor, but not the scaling exponent ν [377]. While Ash1 has a net positive charge (net charge per residue [NCPR] of +0.18), this is not large enough to formally designate the sequence a strong polyelectrolyte, according to previous criterion [126]. Nevertheless, we sought to use limiting models to ask if a polyelectrolyte effect - where like-charged residues repel one another - could account for the conformational behavior of Ash1.

A 1-bead-per-residue coarse-grained lattice model of Ash1 was used (fig 6.5a) (see chapter 14 for model and simulation details). Each residue was defined as being either positively charged (K/R) or uncharged (all other residues). In the simulation, positively charged beads are repulsive for one another over long and short ranges, while uncharged beads are weakly attractive for one another. To modulate the ionic strength, we systematically varied the repulsive charge-charge interaction, from zero charge-charge repulsion (equivalent to infinite ionic strength) upwards. Independent Monte Carlo simulations were run to convergence. As the ionic strength is decreased (and repulsion increased), we observe a systematic expansion

of the protein towards the good solvent limit, defined by a normalized R_G of 1.0 (fig. 6.5b, c). This result is entirely consistent with existing literature on biological polyelectrolytes [405, 425, 546].

Based on SAXS scattering data and all atom simulations, a normalized R_G of 0.85 corresponds to the degree of expansion observed in Ash1 at relatively low NaCl concentration (75 mM). If the expansion of Ash1 is driven by a polyelectrolyte effect, we would expect a statistically significant change in the R_G as a function of NaCl. Salt titrations between 50 mM and 1.5 M indicate no change in expansion. While the models used here represent simplified descriptions (although the same result is recapitulated with much more complex models), it is worth emphasizing that if the origin of expansion in Ash1 can be described by the first order effect of charge repulsion then these models should be sufficient to provide qualitatively accurate predictions regarding the expected behavior as a function of increased salt concentration [425].

In addition to the analysis above, we also compared our results from Ash1 to those from an IDP that does demonstrate expanded behavior due to a polyelectrolyte effect - prothymosin α (ProTα). Unlike Ash1, ProTα would be considered a strong polyelectrolyte with an NCPR of -0.33. At 140 residues with a R_G of $\sim$ 40Å, ProTα is also beyond the dimensions expected for a fully denatured protein of the same length as calculated from the Kohn *et al.* parameters (35.6 Å), making it a convenient case study to compare and contrast with Ash1 [405]. Upon addition of 1 M KCl, the R_G of ProTα decreases to $\sim$30 Å, a result consistent with and explained by polyelectrolyte/polyampholyte theory [405]. Such a collapse is also observed on the addition of 4 M GdmCl for the same reason GdmCl is an ionic denaturant whereas the addition of urea leads to a continuous and modest expansion due to disruption of local structural preferences and/or transient long range interactions.

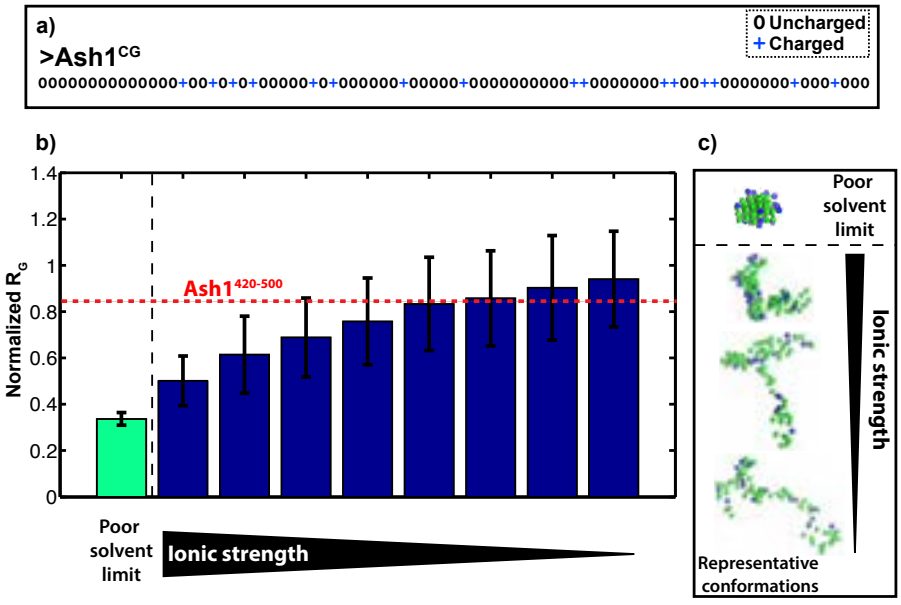

Figure 6.5: (a) The sequence of Ash1 used in the coarse grained simulations (Ash1^{CG}). (b) The normalized R_G associated with the simulations performed at different ionic strengths. The normalized R_G is defined as the R_G from the simulation divided by the R_G of a length-matched polymer behaving in the good solvent limit. This allows us to describe the coarse-grained simulations and all atom simulations in equivalently normalized units. As ionic strength (a mean field description of salt concentration, although we note that ionic strength and salt concentration are not equivalent due to salt-specific effects) increases, there is a concomitant decrease in the R_G [344]. The poor solvent limit is included to highlight the fact that at infinite ionic strength the chain does not undergo complete compaction i.e., we assume some degree of chain expansion mediated by the non-charged residues, but in this limiting model, the primary driver of expansion is charge repulsion. These results are at odds with SAXS results and all atom simulation results of Ash1, where a change in ionic strength between 50 mM to 1.5 M yields no change in the R_G. (c) Representative conformations taken from simulations at different ionic strength.

Condition	R_G **Ash1 (Å)**	R_G **ProT**α **(Å)**
Native Solution (50 mM XCl)	27.4	40.5
High Salt (1 M XCl)	28.6	29.9
Analytical Good Solvent (Kohn *et al.* parameters)	26.3	35.6
Good Solvent (EV simulation)	33.4	45.5
4 M Urea	26.9	43.9
4 M GdmCl	31.0	30.5

Table 6.1: Comparison of solution responsiveness of Ash1 vs. ProTα. XCl used to denote NaCl or KCl. Note that while ProTα shows as 25% compaction in 1 M salt (as expected for a strong polyelectrolyte), Ash1 shows almost no change in global dimensions. A point unrelated to this study, but relevant in the context of chapter 7 is the discrepancy between the predicted good-solvent dimensions of ProTα and simulated EV behaviour. We argue that the 'good solvent' regime defined by Kohn *et al.* is a good solvent regime for foldable proteins (which, in their defence, is exactly how it was defined), but *inherently* captures residual local structure and long-range correlations that suppress the chain's global dimensions with respect to the theoretical limit of a structurally mapped sequence in a *perfect* solvent (uniformly no intra-molecular interactions).

In contrast, upon the addition of 1 M NaCl, the dimensions of Ash1 are unperturbed. The addition of 4 M GmdCl leads to a modest expansion, equivalent to that observed in 4 M urea for ProTα, results that are not explained by polyampholyte and polyelectrolyte theory. These results strongly suggest that while ProTα and Ash1 both occupy highly expanded conformational ensembles, the driving forces leading to this state are fundamentally different.

6.3.5 NMR Reveals Local Changes Upon Phosphorylation

We used NMR spectroscopy to perform comparative assessments of site-specific conformational preferences of Ash1 and pAsh1. Figure 6.6a shows ^{1}H-^{15}N heteronuclear single quantum coherence (HSQC) spectra for Ash1 and pAsh1. The HSQC spectra show poor chemical shift dispersions and sharp line widths for both sequences. This is consistent with averaging of the magnetic environment via interconversion amongst different conformations. Phosphorylation resulted in a downfield shift of ^{1}H resonances and upfield shift of ^{15}N resonances, especially for phosphorylated residues. This is consistent with the presence of phosphorylamide hydrogen bonds [578]. Less pronounced shifts were observed for residues that are proximal in the linear sequence (fig. 6.6). Overall, the NMR data are consistent with specific conformational changes that are localized to phosphorylated residues that accompany multisite phosphorylation of Ash1.

The increase in FCR upon phosphorylation would be expected to promote expansion of pAsh1 when compared to Ash1. The invariance of global conformational properties to multisite phosphorylation suggests the possibility of a compensatory expansion and compaction within the ensemble. Given the possibility of *cis* / *trans* proline isomerization we asked if

an increase in the population of *cis* proline isomers could counter the effects of multisite phosphorylation and explain the invariance of global dimensions.

Due to the repetitive nature of the Ash1 sequence and limited chemical shift dispersion, the proline resonances in ^{1}H-^{13}C HSQC spectra are degenerate, affording a global comparison of *cis* and *trans* proline populations (fig. 6.6b). When averaged over all proline residues in the protein, the *cis* populations in Ash1 and pAsh1 are highly similar. Although the carbon-detect CON spectra showed chemical shift differences between Ash1 and pAsh1, the numbers and intensities of minor proline signals were qualitatively similar (fig. 6.6c).

The low populations of *cis* proline isomers and the relatively low sensitivity of carbon-detect experiments did not allow for the assignment of all minor signals, which are a mixture of *cis* and *trans* proline resonances that are shifted by the sequence proximity of *cis* proline residues (fig. 6.7. Measurements directed at shorter peptides, which recapitulate the sequence-local effects on *cis* / *trans* proline isomerization, confirmed the insensitivity of *cis* proline contents to phosphorylation for each proline within individual phosphosites (fig. 6.8). Our data therefore suggest that the population of cis proline isomers is essentially insensitive to phosphorylation in Ash1. Therefore, changes to the overall content of *cis* proline isomers do not appear to provide compensatory compaction to offset the expected expansion from the increased FCR upon multisite phosphorylation. This is contrast to work on the RNA POL II C-terminal domain (RNA POL II CTD), where phosphorylation leads to a systematic shift from *cis* to *trans*, although we note that in this case of the RNA POL II CTD phosphorylation leads to the sequence changing from a neutral sequence with no charged residues to a strong polyelectrolyte, a fundamentally different transition.

We next asked if multisite phosphorylation leads to a set of new interactions that were not attainable in the unphosphorylated state, and if these interactions might afford compensatory

compaction, assuming an expansion associated with an increase in FCR? One candidate for this type of interaction would be local pSer/Arg salt bridges, which are proposed to lead to compaction within shorter IDRs [309]. We do measure evidence for such local salt bridges in model peptides as indicated by changes in arginine side chain chemical shifts upon phosphorylation (fig. 6.9). However, we do not observe chemical shift differences of Arg sidechains between Ash1 and pAsh1. This suggests that if pSer/Arg salt bridges are present, they are less persistent than in short peptides and cannot provide compensatory compacting effects to offset any expansion derived from the increased FCR due to multisite phosphorylation [199].

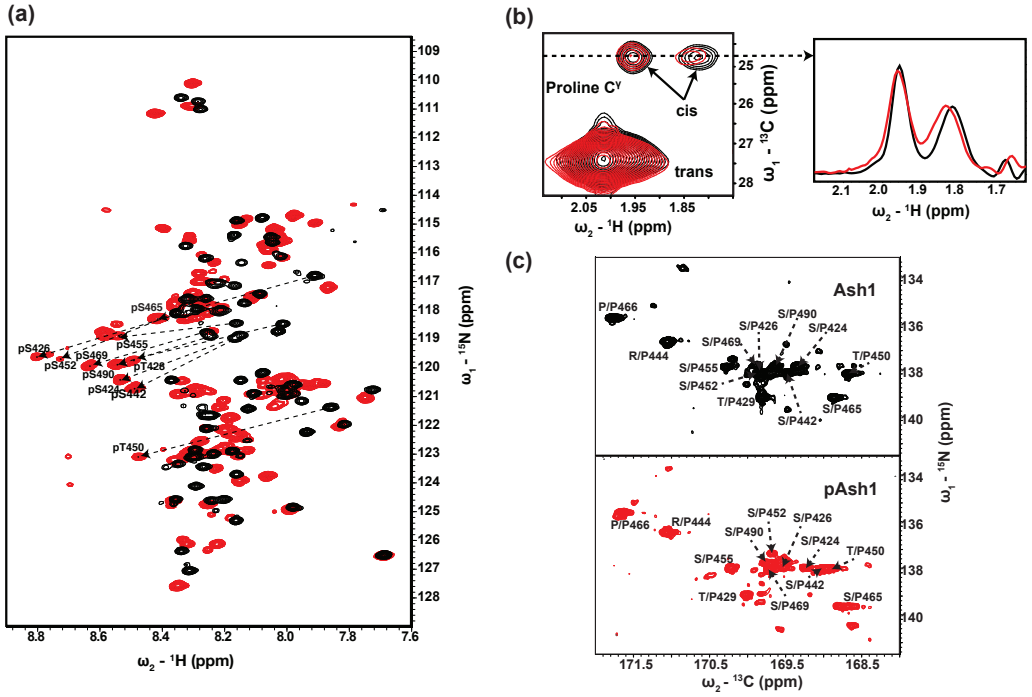

Figure 6.6: (a) Superposition of ^{1}H, ^{15}N HSQC NMR spectra for Ash1 (black) and pAsh1 (red). Signals of phosphorylated residues are labelled and experience a ^{1}H downfield shift at pH 6.95. Additional chemical shift changes indicate local conformational changes. Spectra were fully assigned (see published report for full assignment). (b) Proline C^{γ} region of ^{1}H, ^{13}C HSQC spectra for Ash1 (black) and pAsh1 (red). One strong degenerate resonance from all *trans* proline C is observed, while the *cis* proline C$^{\gamma}$s result in two small signals, with an upfield shift of ∼3 ppm. 1D slices through the *cis* proline signals for Ash1 (black) and pAsh1 (red), intensities normalized to the trans resonances, show similar global *cis* proline populations for Ash1 and pAsh1 of 9.7 ± 1.8 % and 8.4 ± 1.1 %, respectively. (c) Proline region of CON spectra show extensive resonance splitting due to *cis*/*trans* proline isomerization and the extent of splitting is qualitatively similar for Ash1 and pAsh1. The major signals stem from *trans* proline residues, minor signals from *cis* proline residues and trans proline residues in sequence vicinity of a *cis* proline.

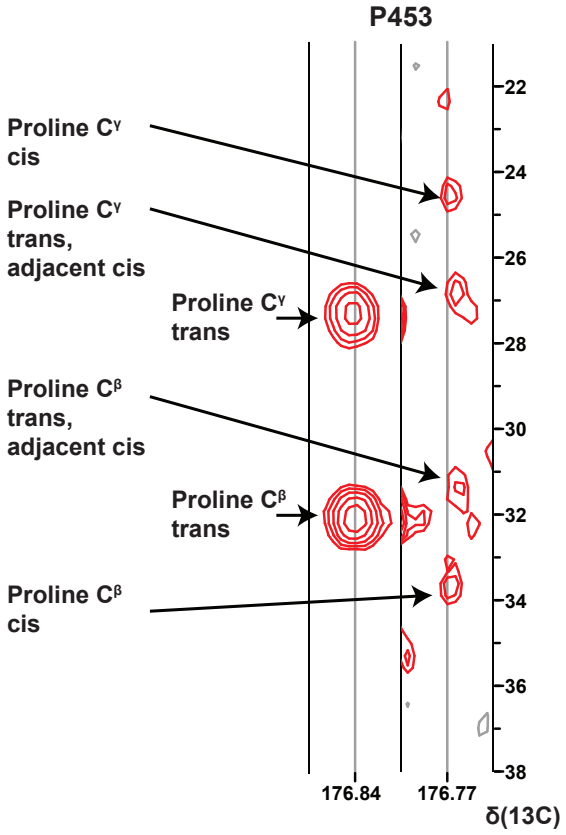

Figure 6.7: The 3D CCCON-IPAP spectrum correlates all sidechain carbons with C'$_i$ (direct dimension) and N$_{i+1}$ (indirect), effectively linking to signals in the 2D CON experiment. *Cis* proline C$_\gamma$ and C$_\beta$ signals are predictably shifted to higher and lower chemical shifts, respectively, as compared to those of trans proline. In theory, minor signals in the 2D should be able to be assigned. However, in practice sidechain resonances from *cis* conformations are only slightly visible above noise level. As an example, the 3D strip for the major *trans* conformation of proline 453 is shown zoomed in on the C$_\gamma$ and C$_\beta$ region. Minor conformations found immediately adjacent in the C'-N plane are shown on the right. The strip on the right contains contributions from P453 in *trans*, but split by an adjacent proline (P451 or P456), and P453 in *cis*. These two species would be unresolved in the 2D CON spectrum and, further, are barely determined above noise levels.

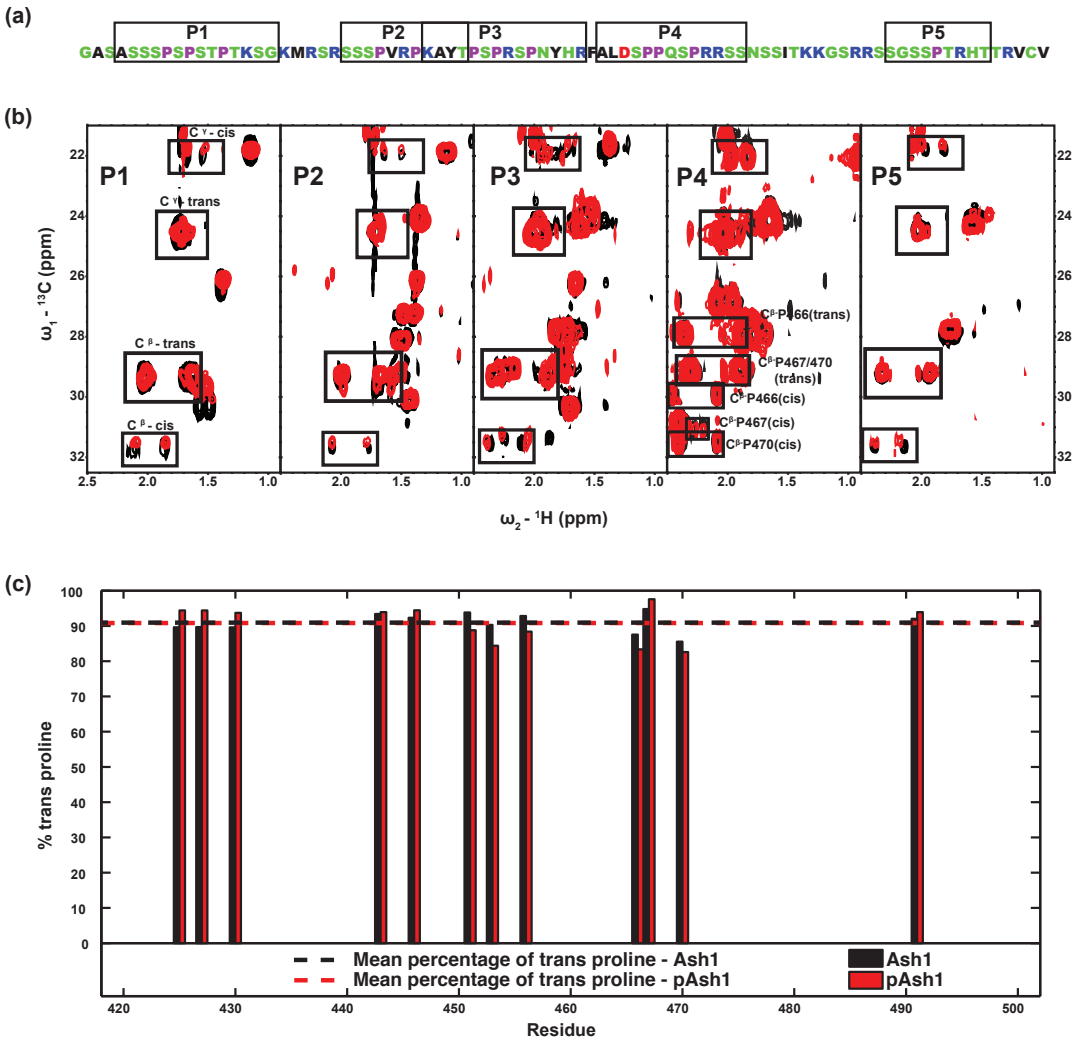

Figure 6.8: (a) The boxes around parts of the Ash1 sequence indicate the boundaries of individual peptides. (b) The C^{γ} and C^{β} containing regions of the natural abundance ^{1}H-^{13}C HSQC spectra for all peptides. Phosphorylated peptides (red) are shown overlaid with the non-phosphorylated (black) counterpart. Visual inspection shows nearly identical *cis*/*trans* proline equilibria. (c) The percentage of *trans* proline residues calculated from individual Ash1 (red) and pAsh1 (black) peptides . For comparison, the mean values are given by dashed lines.

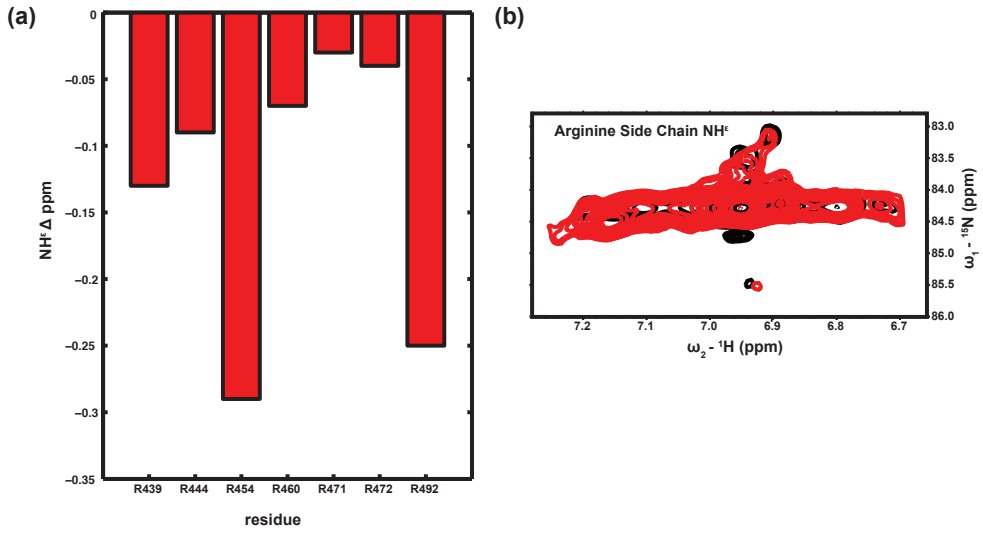

Figure 6.9: (a) The chemical shift perturbations resulting from phosphorylation of arginine NH^{ϵ} protons in individual peptides. The larger perturbations may be indicative of the formation of pSer/Arg salt bridges. (b) The arginine sidechain NH^{ϵ} correlations measured in full length Ash1 (black) and pAsh1 (red). The CP-HISQC pulse sequence, using the default parameter set, was used to obtain the best possible resolution of Arginine sidechains [663]. Additionally, samples were measured at pH 5.8 to minimize signal loss from proton exchange. Measurements at pH 6.95 showed even greater overlap, however this was largely an artefact of poor signal to noise. Ash1 and pAsh1 show nearly identical chemical shifts, showing that pSer/Arg salt bridges, if present, are not stable.

Our results thus far suggest that while phosphorylation has no impact on the global dimensions of Ash1, it does lead to quantifiable local changes, especially in the chemical environments of phosphorylated residues. Our data argue against chain compaction upon multisite phosphorylation due to proline isomerization or persistent pSer/Arg or pThr/Arg salt bridges. The other alternative is that the degeneracy of local / non-local interactions along the chain of a long disordered protein might compete with and compensate the effects of one another. This type of intrachain screening of attractive and repulsive interactions, proposed by Flory, can lead to invariance of global dimensions even after multisite phosphorylation [180]. Such effects are difficult to discern experimentally. Accordingly, we turned to all atom simulations to explore and understand the synergy between sequence-encoded global and local conformational preferences.

6.3.6 All-Atom Simulations Reproduce Ash1 Experimental Results

We performed all atom Metropolis Monte Carlo simulations using the ABSINTH implicit solvation model and forcefield paradigm that we combined with parameters from the OPLS-AA/L molecular mechanics forcefield [613]. These simulations were aided by the development of optimized parameters for proline residues that were made interoperable with the ABSINTH model and OPLS-AA/L forcefield [471]. Solution ions and all polypeptide atoms are modelled explicitly. Parameters for solution ions are interoperable with any solvation paradigm, including ABSINTH [361]. The explicit modelling of solution ions allows us to query the effects of changes to salt concentration on conformational properties. For further discussion on the ABSINTH forcefield see chapter 2.

In the presence of 50 mM NaCl, the ensembles generated by the all atom simulations yield a mean R_G value of 28.9 ± 1.2Å (fig. 6.10a). We obtained similar R_G values from simulations in the presence of 150 mM NaCl. Within experimental error, these values are in agreement with inferences from the SAXS data for Ash1. In order to calibrate the pattern of intra-chain distances in simulation results we generated ensembles to reproduce two theoretical reference limits. These are designated as the Flory Random Coil (FRC) and Excluded Volume (EV) ensembles (see chapter 5). The mean R_G scales with chain length (N) as $N^{0.5}$ and $N^{0.59}$ for FRC and EV ensembles, respectively. We have implemented a method to generate sequence-specific ensembles that conform to the FRC and EV limits. Using this approach, we calculated the mean R_G values for Ash1 in the FRC and EV limits to be 23.4 ± 1.6Å and 33.4 ± 1.8Å, respectively. The mean R_G value and the distribution of R_G values calculated from the ABSINTH ensembles lie in between the FRC and EV limits (fig. 6.10a).

We also performed ABSINTH-based all atom simulations on a phosphomimetic version of Ash1, which we refer to as eAsh1. In these simulations, every phosphorylated Ser/Thr residue of Ash1 was replaced with Glu. The mean R_G value of eAsh1 in the presence of 50 mM NaCl was 27 ± 1.2Å. The mean R_G values for Ash1 and eAsh1 from simulations are within error of one another, and within error of the R_G values determined by SAXS for Ash1 and pAsh1, respectively. The chemical structure of and local conformational properties engendered by Glu and pSer/pThr are distinct from one another. Despite these differences, the agreement between pAsh1 SAXS results and eAsh1 simulation results suggests that the global conformational properties of pAsh1 might be governed by generic features captured by the phosphomimetic eAsh1. We obtained similar R_G values from simulations of eAsh1 in the presence of 150 mM NaCl - an observation that is consistent with the negligible salt dependence observed from SAXS measurements for pAsh1.

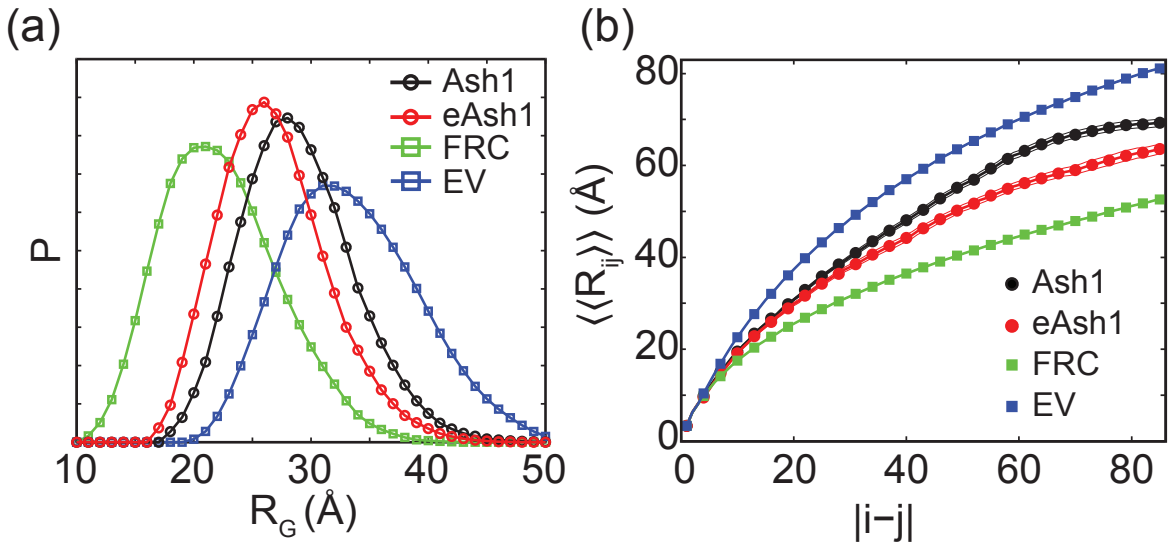

Figure 6.10: (a) The distribution of R_G values obtained from all atom simulations is shown for Ash1, the phosphomimetic eAsh1, and the two reference ensembles, the Flory Random Coil (FRC) and Excluded Volume (EV) limits. Simulations were carried out in the presence of 50 mM NaCl for Ash1 and eAsh1. Ash1 and eAsh1 show similar distributions, with global dimensions between the EV and FRC reference limits. (b) Internal scaling profiles for the four simulations from panel a. For every pair of residues at a given sequence separation ($|i-j|$) the average through-space distance between each pair of residues at that sequence separation, $\langle\langle R_{i,j}\rangle\rangle$, is shown. This provides a summary description of the scaling of intra-chain distances of the polymer. The mean ± SEM is shown as two thinner solid lines.

We used the ABSINTH, FRC and EV ensembles to calculate internal scaling profiles. These profiles quantify the mean spatial separation between all pairs of residues that are $|ji|$ residues apart along the linear sequence (6.10b). The internal scaling analysis represents a formal order parameters in polymer physics theories and is useful for quantifying the intramolecular density of chain atoms around one another and for making quantitative comparisons across different ensembles. A monotonic increase of the spatial separation with sequence separation

is shown by the black and red curves in fig. 6.10b thus confirming the expanded, coil-like nature of the Ash1 and eAsh1 ensembles. The results in fig. 6.10a and 6.10b demonstrate that Ash1 and eAsh1 sample globally similar ensembles that lie between the FRC and EV limits. In addition, simulations of partial phosphomimetic constructs match SAXS results for partially phosphorylated Ash1. Overall, the ABSINTH simulations recapitulate the general insensitivity of global dimensions to changes in the charge states of Ash1.

In order to place the comparison between ABSINTH ensembles and the scattering data on a quantitative footing, we used the ABSINTH all atom ensembles to calculate scattering curves using the CRYSOL package for Ash1 [569]. The results of the comparisons are shown in fig 6.11a (see black curve) and fig. 6.11b. We also calculated scattering curves for eAsh1, and compared those results to the scattering curve from pAsh1 (see red curve in fig 6.11a). In the interest of completeness we also calculated the scattering curves obtained using the Flexible Meccano model (see green curve in fig. 6.11b) [434]. The favourable comparisons between the Flexible Meccano and ABSINTH derived scattering curves as well as between the ABSINTH and experimental data suggest that, on a global scale, the ensembles of Ash1 and eAsh1 resemble that of an expanded random coil whose mean size lies in between two well-defined theoretical limits.

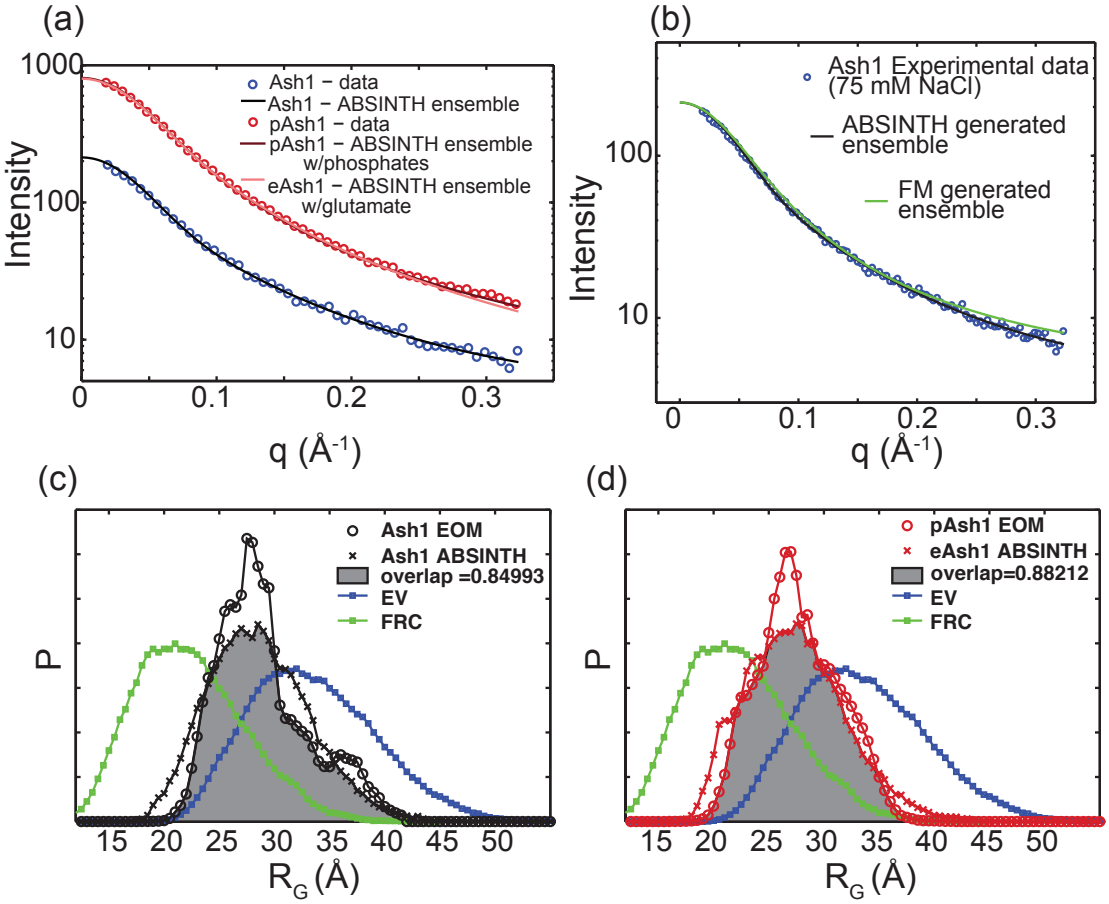

Figure 6.11: (a) Simulation derived scattering curves for Ash1 and eAsh1 compared to the SAXS scattering curve for Ash1 and pAsh1. To generate a pAsh1 ensemble, all phosphomimetic glutamate residues in eAsh1 were substituted by pS or pT residues. (b) Comparison of Ash1 scattering profiles derived from ABSINTH simulations and a Flexible-Meccano ensemble with experimental scattering data. (c) and (d) Overlap of the R_G distributions for Ash1 (c) and e/pAsh1 (d) ensembles generated by EOM or all atom simulations. The overlap is best for the EOM and simulation-derived ensembles. The incomplete overlap is in part caused by the jagged size distributions of the EOM ensembles, which are caused by the fact that they consist of a small number of conformers that collectively agree with the experimental data, but do not explicitly have a coil-like size distribution.

In order to compare the R_G distributions obtained from ABSINTH ensembles and those obtained from the EOM approach, we calculated the degree of overlap between the distributions. Ash1 (SAXS) and Ash1 (simulation) showed a high degree of overlap ($\sim$ 0.85, see fig. 6.11c), highlighting the congruence between simulated ensembles and distributions obtained using models that are designed to match the experimental data. Similarly, the overlap between the EOM and ABSINTH-derived R_G distributions for pAsh1 and eAsh1 is $\sim$ 0.88 (see fig. 6.11d) indicating that the global conformational preferences measured by SAXS for pAsh1 are similar to those obtained from ABSINTH-based simulations of eAsh1. Favourable comparison between the measured scattering curve of pAsh1 and the calculated scattering curve of eAsh1 suggests that the phosphomimetic sequence captures the global conformational preferences of pAsh1. Accordingly, a detailed analysis of these ensembles should provide an explanation for the observed coil-like conformations of Ash1 and the invariance of global conformational properties to multisite phosphorylation.

6.3.7 Sequence Determinants of Ash1 Expansion

We examined sequence features of Ash1 to uncover the source of the intrinsic, sequence-encoded expansion. Ash1 has a proline content of 15%. This is relevant because published heuristics regarding composition-to-conformation relationships of IDRs were derived from simulation results and spectroscopic investigations of sequences with low proline contents.20 In Ash1, 35% of the residues are either proline or charged. Since proline and charged residues respectively drive local and global expansion, we reasoned that the linear sequence distribution of proline and charged residues might explain the observed expansion of Ash1.

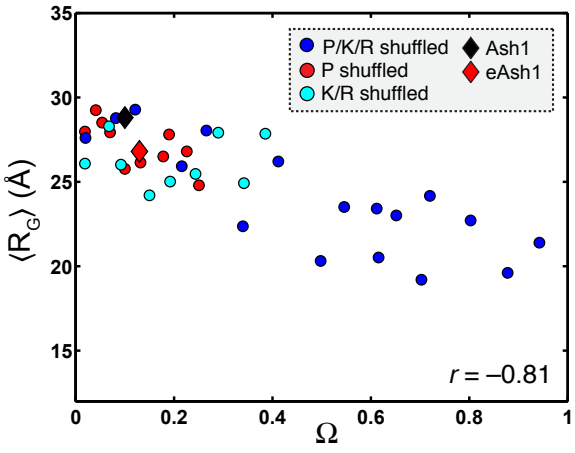

Figure 6.12: Summary of simulation results showing the variation of radii of gyration with Ω. A rational sequence design algorithm was deployed to generate 30 distinct sequence permutants by changing the patterning of proline and charged residues with respect to all other residues. This was achieved by shuffling the positions of proline and charged residues and fixing the positions of all other residues. The complete set of sequences can be found in fig 6.13. Three independent ABSINTH simulations were run for each permutant to determine the mean radius of gyration associated with the ensemble. The results show an inverse correlation between Ω and the R_G.

In Ash1, the proline and charged residues are uniformly distributed with respect to all other residue types along the linear sequence (Figure 1). We quantified this as the mixing or segregation of proline and charged residues (Pro, Lys, Arg, Asp, Glu) with respect to all other residues (Xaa). Specifically, we computed a normalized patterning parameter designated as Ω where $0\ \Omega\ 1$. Our definition of Ω is analogous to the definition of the parameter introduced by Das and Pappu to quantify the mixing vs. segregation of oppositely charged residues [126]. The calculation of Ω is described in the methods section. If proline and charged residues are well mixed with respect to all other residues, then the value of Ω for the sequence of interest approaches zero. Conversely, if proline and charged residues are segregated with respect to all other residues in the sequence of interest, then Ω approaches 1.0. We find that $\Omega = 0.1$ for Ash1 and 0.13 for pAsh1/eAsh1. Therefore, we hypothesized that the uniform distribution of expansion-driving proline and charged residues along the Ash1 / pAsh1 / eAsh1 sequences give rise to a sequence-encoded preference for expanded conformations.

To test this hypothesis, we used an unbiased sequence design algorithm to design a series of sequence permutants of Ash1. These permutants - all of which have an identical amino acid composition - were generated by shuffling the positions of proline residues (red symbols in fig 6.12), charged residues (cyan), or both (dark blue). Using this approach we generated sequences corresponding to different values of Ω. The complete set of sequences can be found in fig. 6.13. We performed multiple independent atomistic simulations for each of the Ω-permutants. Figure 6.12 shows the calculated R_G for each Ω-permutant plotted against Ω. This analysis shows a strong negative correlation (Pearson's correlation coefficient $= -0.81$) between the degree of expansion and Ω, suggesting that the mixing or segregation of proline and charged residues with respect to other residues engenders expansion versus compaction, respectively. This analysis provides a plausible explanation for the sequence-encoded preference for expanded Ash1 ensembles in aqueous solvents.

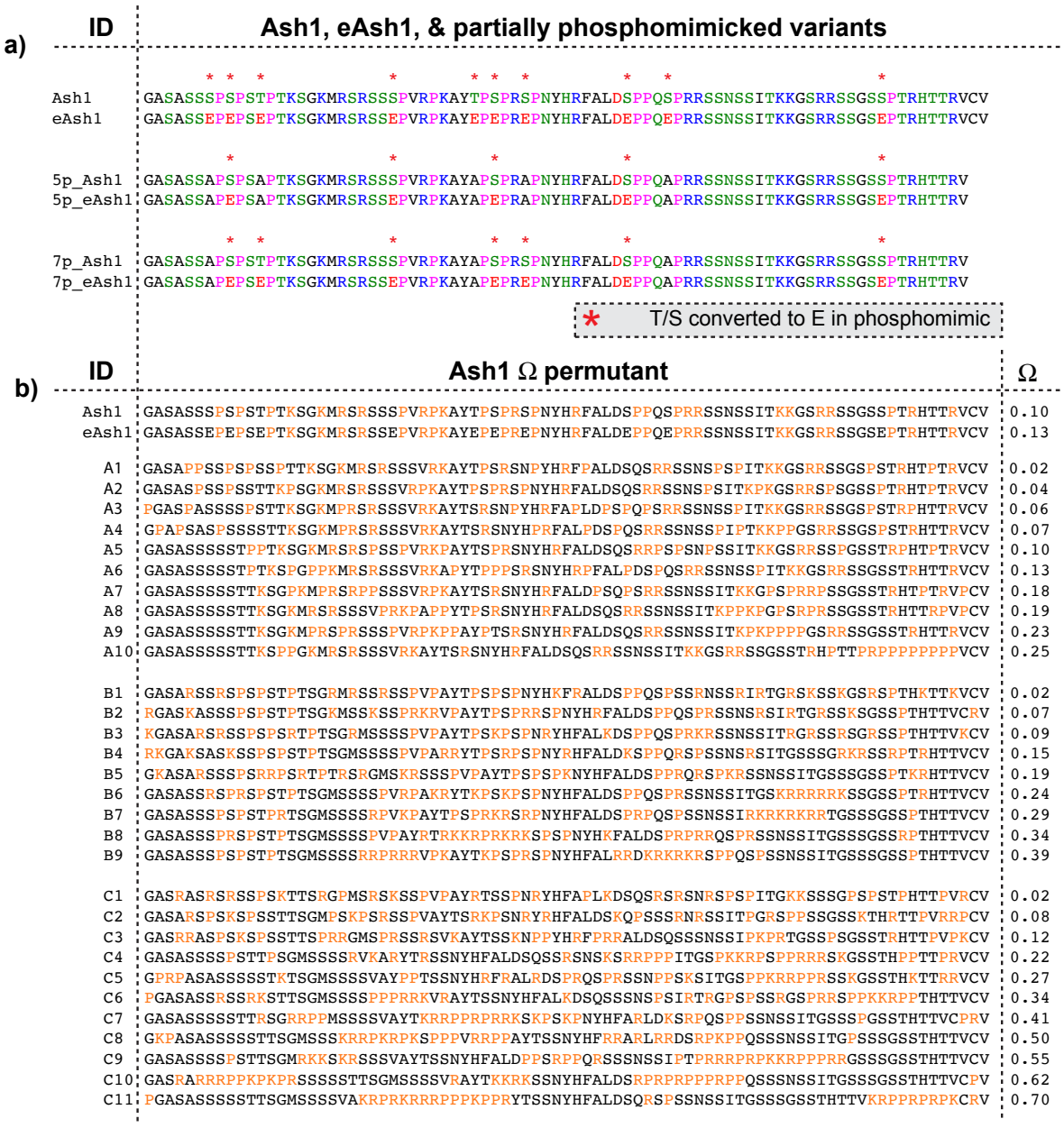

Figure 6.13: (a) The sequences of Ash1 and eAsh1 compared to the two permutants with 3 (7p Ash1) and 5 (5p Ash1) phosphorylation sites replaced by alanine. (b) All sequence variants designed using an unbiased algorithm to properly sample Ω.

6.3.8 The Compensatory Conformational Changes Allow Global Invariance

We quantified secondary structure propensities by comparing NMR derived Cα, Cβ, C' and N chemical shifts with random coil values. We used three different methods (SSP, ncSCP, and δ2D) to convert the measured chemical shifts to estimates of local structural propensities [85, 363, 570]. Although there is reasonable agreement among the estimates obtained using the three methods, there is also considerable variation suggesting the need for caution in extracting precise quantitative trends from the experimental data. We also used the simulated ensembles for Ash1 and eAsh1 to calculate local structural propensities. In order to avoid comparisons among metadata, we compared the results from our analysis of experimental data to analysis of simulation results that are based on backbone / angles as implemented in the BBSEG algorithm that is part of the CAMPARI modelling suite (fig 6.15).

Overall, the local structural propensities calculated from simulations agree with the consensus interpretation that emerges from analysis of experimental data. All four methods point to an increase in α-helical propensities upon phosphorylation. This increase in helicity is around residue 430-435 and residue 470-480.

We also examined the propensities for polyproline II (PPII) conformations (fig. 6.13). Although there are modest changes in PPII propensities upon phosphorylation, both ensembles appeared to have a relatively high PPII propensity across the entire sequence. Analysis of the simulation results using BBSEG and of the experimental data using δ2D allow us to evaluate the PPII propensity for each residue. We also evaluated the simulated ensembles for persistent PPII preferences across consecutive stretches along the linear sequence. This

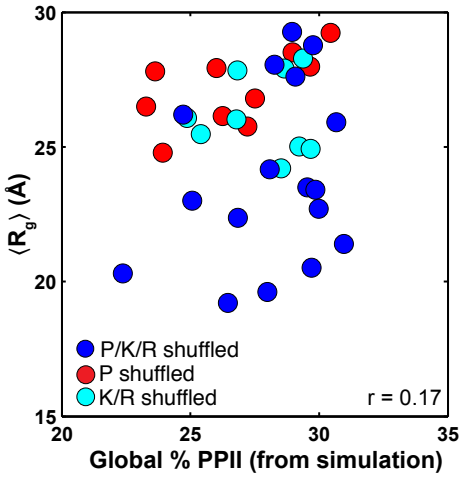

Figure 6.14: R_G versus the fraction of PPII content for Ω permutants. The global fraction of PPII conformations was calculated from simulations of all Ω permutants listed in fig. 6.5. There is no correlation between increased PPII propensities and R_G values

analysis suggests that while individual residues have distinct preferences for the PPII basin of Ramachandran space, these local preferences derive from uncorrelated transitions into and out of the PPII basin. Accordingly, the expansion of Ash1 cannot be attributed to persistent preference for PPII helices, which require that at least three consecutive residues simultaneously occupy the PPII basin. Instead, the overall expansion of Ash1 can be attributed to the synergistic combination of proline and charge contents, the uniform mixing of these residues, the local stiffening due to proline residues, and the favourable solvation of charged residues. To further explore the relationship between PPII and conformational behaviour we examined the correlation between total fractional PPII behaviour and global dimensions. We found no correlation between global PPII occupancy and the R_G values for the series of Ω permutants of Ash1 (fig. 6.14), suggesting that sequences with similar PPII propensities can have very different R_G values.

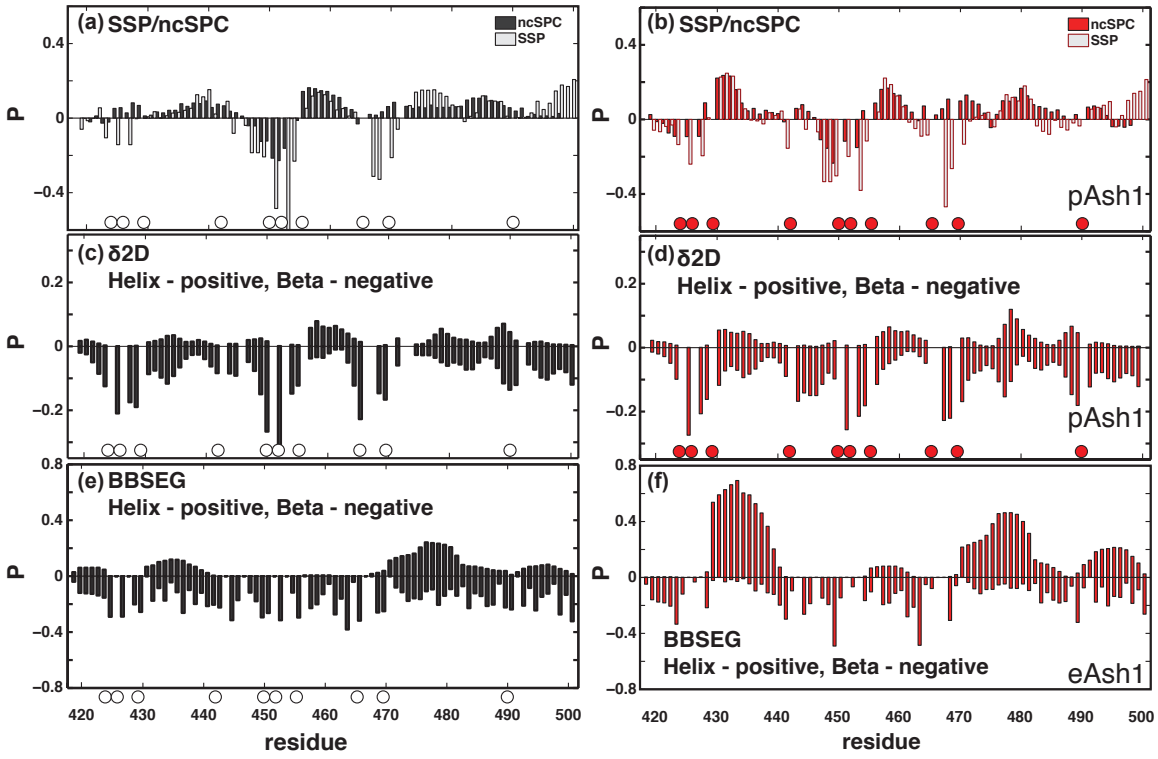

Figure 6.15: Secondary structure propensities from chemical shift data and simulation. The circles define the position of phosphosites. (a, b) Secondary structure propensities for Ash1 (black) and pAsh1 (red) calculated from Cα, Cβ, and C' chemical shifts using SSP and ncSPC [363, 570]. (c, d) Secondary structure propensities for Ash1 (black) and pAsh1 (red) calculated from Cα, Cβ, C' and N chemical shifts using δ2D [85]. (e, f) Secondary structure propensities for Ash1 (black) and pAsh1 (red) calculated from atomistic simulations using the distributions of backbone dihedral angles (BBSEG).

We next asked if the invariance of global conformational properties between Ash1 and pAsh1 / eAsh1 might derive from compensatory changes in the patterns of preferred intramolecular distances. Using the ABSINTH-based ensembles for Ash1 and eAsh1 we calculated the ensemble-averaged distances between the centers-of-mass of every unique pair of residues in the sequence. Figure 6.16a shows the raw data with the upper triangular portion corresponding to Ash1 and the lower triangular portion corresponding to eAsh1. Given the coil-like nature and the wide range of inter-residue distances within the ensembles for both sequences, it is difficult to uncover the important distinctions between the two ensembles. This is remedied by calculating normalized distances, whereby the distance for every pair of residues is normalized by the value we obtain for the sequences in the EV limit. These two-dimensional scaling maps are shown in fig. 6.16b. The scaling maps reveal the following insights: Residues 455-460 in eAsh1 make long-range contacts with spatial separations in the range of 25 - 35 Å with residues 474-490. Residues 455-460 contain four phosphosites and two Arg residues, whereas the region spanning 474-490 contains eight Lys/Arg residues and one phosphosite. This suggests the presence of non-local, intermediate-range electrostatic interactions between a cluster of positively charged residues near the C-terminal region and a cluster of negatively charged residues in the central region. These complementary, non-local electrostatic interactions engender a modest compaction with respect to the EV limit that is not observed in the Ash1 ensemble. However, the effects of compaction are offset by expansion vis-a-vis the EV limit across the region spanning residues 450-470. An explicit example of the local compensatory changes is shown in fig. 6.16c, where the dimensions of two sub-peptides examined in the context of Ash1 and eAsh1 are compared. We attribute this expansion to enhanced electrostatic repulsions and the favourable free energy of solvation of the negatively charged residues within this region. Finally, we observe a modest local

compaction and longer-range expansion for the region spanning residues 435-440, which is attributable to the increased α-helix propensity upon phosphorylation (figs. 6.15 and 6.16b).

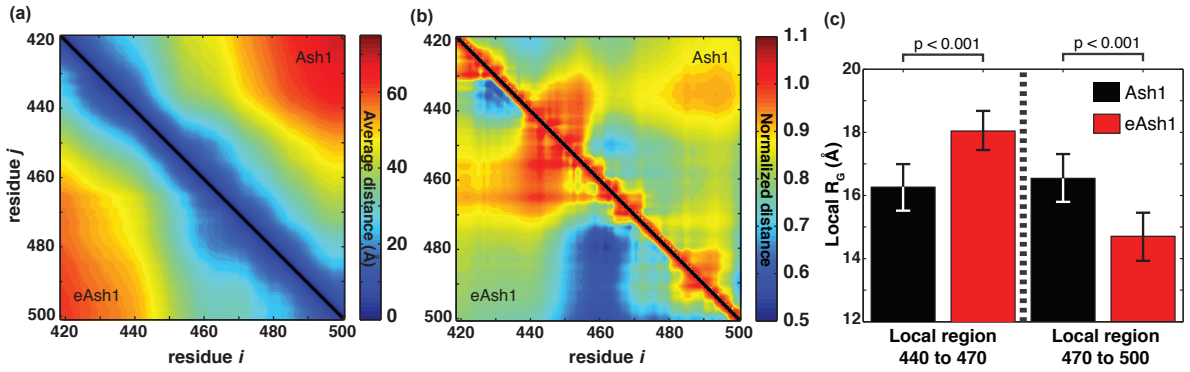

Figure 6.16: (a) The distance map summarizes the average distance between each pair of residues in Ash1 (upper triangle) and eAsh1 (lower triangle). Both Ash1 and eAsh1 show apparently uniform expansion across all length scales, consistent with expanded, coil-like ensembles sampled by both sequences. (b) All inter-residue values in panel (a) were normalized using the inter-residue distances from an EV simulation for Ash1 (upper triangle) and eAsh1 (lower triangle). This leads to a normalized scaling map. Regional biases for compaction (scaled distances less than unity) or expansion (scaled distances greater than unity) become clearer when operating in a normalized distance space. (c) We calculated the local R_G associated with two sub-peptides in the context of the full chain to demonstrate the compensatory changes observed in Ash1 vs. eAsh1. The 30-residue stretch between residue 440 and 470 is more expanded in eAsh1 than in Ash1, while the 30-residue stretch between residues 470 and 500 is more compact. These regions were identified from the scaling maps in panel (b).

Importantly, the changes observed within the eAsh1 ensemble are mutually compensatory. This is an example of intra-chain screening that is a central tenet of Flory's theory for realizing unperturbed global dimensions [180]. Repulsive interactions that lead to local / non-local chain expansion are screened by the effects of attractive interactions that lead to local / non-local chain contraction. Since these compensatory interactions involve partially overlapping regions of the sequence, the intra-chain screening leads to unperturbed chain dimensions when compared to the unphosphorylated ensemble. Additionally, the negligible salt dependence of the phosphorylated ensemble is explained by weak screening provided by solution ions when compared to the screening of repulsive interactions by attractive ones that are encoded by the sequence, which also controls the effects of post-translational modifications.

Compensatory changes in conformational dynamics are amenable to scrutiny via NMR relaxation methods. Figure 6.17a and b show a comparative analysis of the spin-lattice relaxation rates (R_1) and the spin-spin relaxation rates (R_2) for Ash1 versus pAsh1. While R_1 rates and heteronuclear NOE values (fig. 6.17c and d) are similar for both Ash1 and pAsh1, there are discernible jumps in R_2 rates in clusters along the pAsh1 sequence [293]. These enhanced R_2 rates are indicative of a slowdown in local dynamics upon phosphorylation caused by transient interactions, in agreement with the proposed model of competing local / nonlocal interactions. Specifically, enhanced R_2 rates in the central region ($\sim$450-460) and less pronounced clusters toward the C-terminus are consistent with the main regions identified from analysis of simulation results as being involved in long-range electrostatic interactions upon multisite phosphorylation.

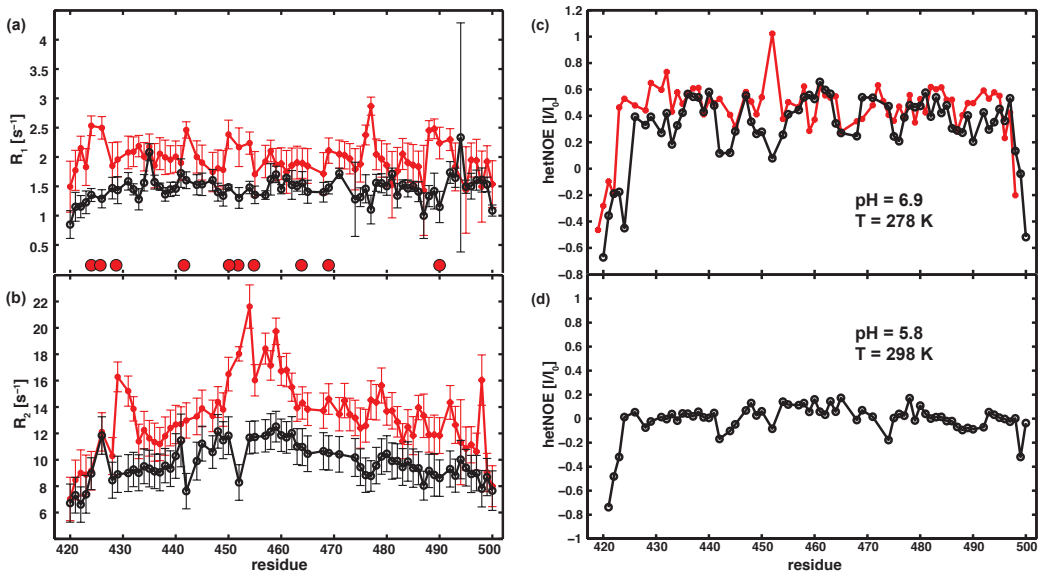

Figure 6.17: (a) Ash1 (black) and pAsh1 (red) R_1 rates and (b) R_2 rates. Phosphorylation sites are marked by red circles. Enhanced R_2 rates for pAsh1 are in agreement with competing transient interactions. (c) Heteronuclear NOE values were collected with a 3.5 second relaxation delay and with and without a 5 second presaturation delay. NOE values for Ash1 (black) and pAsh1 (red) recorded at 278 K. (b) NOE values for Ash1 recorded at 298 K. The pH was lowered to 5.8 to minimize loss of signal due to solvent exchange. Low temperature NOE values are nearly uniformly high. For Ash1, values are reduced to near zero or negative at room temperature.

6.4 Discussion

6.4.1 Context is a Crucial Modulator of Conformational Behaviour

Overall, the simulation results provide a nuanced description of how multisite phosphorylation might influence the conformational properties of Ash1. The effects of local / non-local expansion and compaction involving partially overlapping sequence regions leads to

unperturbed global dimensions with respect to the unphosphorylated Ash1. This Flory-like screening of intra-chain attractions by repulsions is encoded by the amino acid sequence of Ash1, which controls the overall conformational properties prior to and upon multisite phosphorylation. With regard to the latter, it is worth noting that the patterning of proline and charged residues with respect to all other residues changes only slightly upon multisite phosphorylation. This is quantified in terms of the value of Ω which changes from 0.1 to 0.13 (fig. 6.11a) implying a uniform dispersion of proline and charged residues along both sequences.

Our NMR data do not directly report on weak, transient compensatory local / non-local interactions, which seem to be the driving forces of the expanded global dimensions of Ash1 and pAsh1. However, the lack of observable stable structural motifs such as persistent salt bridges, and the highly averaged chemical shifts are consistent with transient, competing interactions in Ash1 and pAsh1. In long IDRs, the balance of local / non-local interactions strongly depends on their patterning along the sequence and this determines whether interactions spanning distinct spatial scales reinforce or compete with each other. In shorter peptides, the competition from truly long-range interactions is absent. Hence, the effect of local interactions on the global conformational properties will be more direct. Although the 81 residue stretch is significantly larger than the 15-30 residue fragments often examined, even in Ash1 there remains a substantial sequence context that we have ignored. Therefore, understanding the hierarchical influence of sequence and structural contexts on the conformational properties of IDRs remains an open challenge.

6.4.2 Sequence Features of Ash1 are Shared by Other IDRs

We asked if the patterning of proline and charged residues with respect to other residues is a feature that is shared by other proteins that undergo multisite phosphorylation. A conservative search through the human proteome for proline-rich regions that are predicted to be disordered and undergo multisite phosphorylation identified a number of putative candidate regions. For the full list please see the table included in the supplementary information associated with the published paper.

The proline-rich region of the microtubule-associated protein tau shares many of the sequence features of Ash1; in a 90-residue stretch, it contains 13 phosphorylation sites and 22 proline residues. A recent Förster Resonance Energy Transfer study of a 14 residue peptide extracted from this region demonstrated an expansion upon phosphorylation [97]. Earlier studies showed that multisite phosphorylation causes local conformational changes, as determined by NMR, while global dimensions measured by SAXS remain unperturbed in a phosphomimetic construct [411, 525]. These observations - global insensitivity and local changes - are highly reminiscent of our results from Ash1. The *S. cerevisiae* cyclin-dependent kinase (CDK) inhibitor Sic1 undergoes multisite phosphorylation, triggering its degradation and subsequent cell cycle progression [531, 611]. The overall dimensions of the non-phosphorylated and phosphorylated states are highly similar as determined by SAXS and NMR, and yet there are extensive local conformational changes [392, 393].

A relatively uncharacterized protein, Chromosome alignment-maintaining phosphoprotein (CHAMP1), contains a 350 residue, proline-rich domain that undergoes extensive CDK1-dependent phosphorylation during mitosis. Given the results for Ash1, we may expect this

proline-rich region to form a highly expanded ensemble, with the overall dimensions remaining unperturbed in response to varying degrees of phosphorylation. This protein may act as an electrostatically-tunable scaffold, whereby the degree of phosphorylation influences intermolecular repulsion without significantly altering intramolecular interactions, similar in spirit to the proposed mechanism associated with neurofilament sidearms [311].

We also found a number of disordered regions that undergo multisite phosphorylation that do not have the sequence characteristics of Ash1. These sequences are deficient in proline residues and they are weak / strong polyampholytes rather than polyelectrolytes. How these different regions respond to multisite phosphorylation will depend on their specific sequence contexts and the patterning of relevant residues therein. A key question is if all IDRs that undergo multisite phosphorylation will show a global conformational insensitivity to phosphorylation? Clearly, in some proteins, specific local / non-local interactions form efficiently, because of a lack of competing interactions along the chain. As an example, the protein 4E-BP2, which regulates the initiation of cap-dependent mRNA translation, folds into a stable structure upon multisite phosphorylation that is able to form a complex with its binding partner eIF4E [23]. In this case, multisite phosphorylation generates synergistic, long-range conformational changes. These must be encoded in the sequence as well, albeit by different sequence features.

Our findings for Ash1 and pAsh1 lead to the proposal of a synergistic relationship between proline and charged amino acids that results in expanded conformations of a disordered protein that undergoes multisite phosphorylation, irrespective of its phosphorylation state. Importantly, proline residues appear to offer a mode of local expansion that is independent of the charged residues a property that may be desirable in regions that undergo reversible

changes in local charge density mediated by phosphorylation. IDRs that undergo multisite phosphorylation may in general utilize such proline-based conformational buffering to provide access to modifying enzymes and downstream signalling effectors. Further work is needed to determine the connections between sequence encoded global and local conformational properties and the functional consequences for IDRs prior to and upon multisite phosphorylation.

Chapter 7

Exploring the Unfolded State Under Folding Conditions

The following section is taken from a manuscript with the working title **Direct Observation of a Protein Folding Contraction and Evaluation of Unfolded State Properties using Non-Invasive Time-Resolved FRET** by I. Peran*, A.S. Holehouse*, R.V. Pappu, I.S. Carrico, O. Bilsel, and D.P. Raleigh (*denotes co-first authors). All experimental work (and associated data analysis) discussed in this chapter was performed by I.P. and O.B. A.S.H. performed all simulation analysis. Proteins were prepared and characterized by I.P in the laboratory of D.P.R at the Stony Brook University and rapid kinetic measurements were made in the laboratory of O. B. at the University Massachusetts Medical School by I.P. and O.B.

7.1 Introduction

Despite being the focus of intensive study for over forty years, many questions surrounding protein folding remain unanswered. Recently, there has been substantial progress in characterizing folding pathways using a combination of novel experimental approaches and improved computational methods [105,339,617]. However, the initial stages of protein folding remain unclear, even for simple single domain globular proteins which undergo an apparent two-state folding transition [219, 255, 352, 557, 661].

The unfolded state under highly denaturing conditions is well described by the infinite chain limit for a polymer in a good solvent - a self-avoiding random walk. It is highly expanded with reduced local and long-range interactions when compared to the native state [297, 591, 641]. In contrast, the unfolded state under folding conditions remains less well characterized. Folded globular proteins are typically compact, and their global dimensions are reasonably well described as a polymer in a poor solvent [144]. Given these observations, upon dilution from high concentrations of denaturant into native conditions, an expanded and unfolded globular protein must undergo a collapse transition to progress to its compact native state. A rapid collapse upon dilution has been observed in refolding experiments for some, but not all proteins [255, 352, 661]. Determining the position of the collapse transition along the folding reaction coordinate and characterizing the conformational propensities of unfolded ensembles prior to and following this collapse are essential steps towards deciphering the folding mechanism, and for developing a complete understanding how the amino acid sequence determines the solution behaviour of proteins.

If collapse precedes folding then the ensembles populated following collapse but prior to the folding transition will be most representative of the unfolded state populated under native

conditions. These states are more relevant to protein folding in a cellular setting than the high denatured unfolded state, and provide an appropriate reference state for thermodynamics measurements. The properties of the unfolded state also influence the tendency of proteins to aggregate, with important implications for protein design and human disease [160, 259]. While our focus here is on the unfolded states of foldable proteins, numerous functional proteins are constitutively unstructured under native conditions. For these intrinsically disordered proteins (IDPs), a better understanding of the unfolded states under native conditions of foldable proteins provides an orthogonal set of insights to enhance our understanding of the relationship between primary sequence and conformational ensemble [127, 158, 604].

Perturbing conditions such as high concentrations of denaturant, extreme temperature, and extreme pH, are widely employed to populate the unfolded state so that it can be studied at equilibrium. Expanded unfolded states are sampled in high concentrations of denaturant due to favourable interactions between the protein backbone and sidechains with denaturant [60, 297, 387, 592]. Despite this expanded behaviour, transient long range contacts can form in the urea unfolded state at levels significantly higher than expected for a polymer in a truly good solvent [293, 377]. The formation of long-range contacts in the unfolded state are fully compatible with random coil scaling laws and with radii of gyration (R_G) that are expected for highly unfolded states. It remains unknown if the same contacts are formed in unfolded states in the absence of denaturant.

By definition, the unfolded state under native conditions is transient the free energy balance strongly favours the folded state and protein folding is usually highly cooperative. Consequently, the study of this transient state requires high time resolution measurements

to 'catch; the unfolded state before folding has occurred. The combination of rapid mixing techniques with spectroscopic measurements has allowed the direct interrogation of unfolded states under near native conditions in the absence of strongly destabilizing mutations [48, 492, 655]. SAXS integrated with stopped-flow or microfluidic mixing has also been used to study the early stages of protein folding upon dilution out of high concentrations of denaturant. SAXS experiments have often - but not always - failed to detect collapse prior to the folding transition for two-state folding proteins of < 150 residues [12, 255, 291, 661]. FRET experiments, also in combination with stopped-flow or microfluidic techniques, suggest that the collapse of the unfolded state occurs rapidly and prior to the folding transition [11, 20, 59, 352, 509]. Single-molecule fluorescence experiments have also revealed that some proteins exhibit a continuous contraction of the unfolded state as the concentration of denaturant is decreased [136, 234, 244, 522, 538, 577, 673].

Collectively, these experiments yield conflicting views of chain compaction and of the unfolded state under native conditions. This may reflect the fact that different proteins behave differently, but there may also be a contribution from the inherent limitations of each method. SAXS provides global information about R_G and, with high enough signal-to-noise data, information about overall shape. However, it also offers limited structural resolution, requires high protein concentrations for adequate signal to noise, is more sensitive to expanded conformations, and can be insensitive to transient long range contacts [347,377]. FRET provides specific pairwise distance distributions and can be performed at lower protein concentrations. Although it is potentially more sensitive to the presence of long-range contacts and compact conformations, ensemble averaged and single molecule FRET studies make use of bright dyes, which are invariably large and which have large R_0 values. Appending large aromatic dyes connected by flexible linkers can perturb the system, especially for small single domain proteins, while the inherent large R_0 (see chapter 2 for further discussion on FRET) means

that the experiment can be less sensitive to the conformational transitions which occur on the length scale associated with collapse in small single domain proteins. In addition, extraction of pairwise distances and global properties from FRET data depends on an underlying model to translate transfer efficiencies into distances, which can introduce biases into the interpreted distances [428, 552].

To obtain a high-resolution description of the unfolded state under native conditions we combined time resolved FRET with time resolved SAXS, extensive simulations, and polymer theory. The N-terminal domain of the ribosomal protein L9 1-56 (NTL9) was used as a model systems, as it shows well defined two-state folding behaviour and has been extensively in the context of protein folding [10, 98, 233, 308, 376, 377, 617]. We monitored chain collapse in real time using a sensitive and non-perturbing FRET method that exploits *p*-cyanophenylalanine (FCN) and Trp pairs. FCN is the cyano analog of Tyr and, unlike large dyes traditionally used, represents a minimally perturbing substitution. FCN acts as the donor to Trp and the R_0 is 16 Å [597]. The residue can be incorporated into proteins using solid phase peptide synthesis or recombinantly using the 21st pair technology of Schultz and Mehl [395,626]. FCN fluorescence can be excited selectively in the presence of Trp and Tyr and the fluorescence decay of FCN is single-exponential, facilitating the analysis of time-resolved studies.

We used time-resolved FRET to measure multiple pairwise distance distributions for the unfolded state populated under strongly denaturing conditions (10 M urea) and under native conditions (1 M urea). Continuous-flow methods interfaced with time-resolved detection were used to measure pairwise distance distributions for the unfolded protein in 1 M urea, after dilution out of high denaturant. The FRET studies were complimented by continuous-flow SAXS measurements. The protein folding time is on the order of 2.5 ms, while the FRET and SAXS measurement dead times are 85 µs and 200 µs, respectively. Finally, a globally

consistent ensemble of 60,000 conformations was generated using all atom simulations. The data reveals that for NTL9, modest contraction occurs rapidly, on a time-scale faster than folding. The global ensemble-average dimensions of the chain are representative of a flexible polymer in a Θ solvent, but the chain contains fluctuating native and non-native elements of structure which are more persistent than the contacts sampled in high concentrations of urea. The study highlights the power of non-perturbing fluorescence probes for following rapid conformational changes, and the importance of combining multiple pair positions to construct an accurate, global description of the conformational behaviour.

7.2 Methods

7.2.1 Protein Expression and Purification

p-Cyanophenylalanine (FCN) was incorporated into NTL9 using 21st pair technology developed by the Schultz and Mehl labs [395, 626]. A copy of the wild-type NTL9 gene in the pBAD plasmid (the empty pBAD plasmid was obtained from Prof. Ryan Mehl) was used for mutagenesis. The codon for Tyr 25 was mutated to either Phe (donor only, to avoid undesirable FRET between FCN and Tyr) or Trp (donor/acceptor). The codon for the position at which FCN was introduced was mutated to TAG. The following residues were mutated to FCN: K2, K10, Q33, N42, and N43. There were two additional K2FCN constructs for which the acceptor was placed at either position 33 or 51 instead of at Y25 by mutating Gln or Lys, respectively. For these constructs, Tyr 25 was mutated to Phe to avoid any undesirable FRET with Tyr. In total, there are seven donor/acceptor constructs: Q33FCN-Y25W, K10FCN-Y25W, N42FCN-Y25W, N43FCN-Y25W, K2FCN-Y25W, K2FCN-Y25F-Q33W and K2FCN-Y25F-K51W.

The pBAD plasmid carrying the NTL9 gene along with the pDule plasmid (also obtained from Prof. Mehl), which encodes the aminoacyl tRNA/tRNA synthetase pair, were co-transformed into BL21-AI cells. 4 ml of overnight cell culture was added to arabinose auto-induction media that was prepared as described in J.T. Hammill et. al3. FCN was dissolved in 18 MΩ H2O for a final concentration of 125 mM (NaOH was added for a final concentration of 160 mM to fully dissolve the amino acid). Dissolved FCN was quickly added to media for a final concentration of 1 mM and cells were incubated at 37 °C for approximately 24 hours while shaking. Cells were then pelleted by centrifugation at 5,000 rpm for 15 min

and the pellet was stored at -80°C until purification. Wild-type NTL9 used for small-angle X-ray scattering studies was expressed as previously described4. For purification, cell pellets were suspended in 20 mM Tris buffer at pH 7.5 and lysed by sonication or using a Constant Systems cell disrupter. Protein was purified from the supernatant by cation-exchange chromatography, followed by reverse-phase HPLC on a Vydac C8 or C18 preparative column. For purification with HPLC, an A-B gradient system was used in which buffer A consisted of 0.1% (v/v) solution of trifluoroacetic acid (TFA) in water, and buffer B consisted of 90% (v/v) acetonitrile, 10% (v/v) water, and 0.1% (v/v) TFA. The expression yield was between 10 mg and 30 mg for all mutants, which is 10-40% of the wild-type yield. The purity of each construct was checked using analytical HPLC with a Vydac C18 analytical column. Matrix-assisted laser desorption ionization time-of-flight (MALDI-TOF) mass spectrometry was used to confirm the molecular weight.

7.2.2 Variant Stability

Seven distinct variants of NTL9 containing FCN and Trp were designed and generated. Matrix-assisted laser desorption ionization time-of-flight (MALDI-TOF) mass spectrometry was used to confirm the molecular weight of each species. Variants containing the donor only and the donor in the presence of acceptor were prepared recombinantly using 21st pair technology developed by the Schultz and Mehl labs. All the mutants are folded as judged by CD, are not significantly destabilized compared to the wild-type protein, and all display sigmoidal unfolding transitions (fig. 7.1). Only one of the variants, N43FCN, is destabilized by > 1 kcal/mol relative to wild-type. However, the CD spectrum of this mutant is very similar to that of wild-type NTL9, indicating that overall native secondary structure is not perturbed. Observed relaxation rates in 1 M urea were determined for the variants from

continuous-flow experiments and do not deviate significantly from that of the wild-type protein.

CD spectra were collected on an Applied Photophysics Chirascan instrument and on an AVIV CD spectrometer. Lyopholized protein was dissolved in 20 mM sodium acetate buffer with 100 mM NaCl at pH 5.5 The concentration of protein was 15 - 25 µM. Spectra were recorded at 25°C. CD-monitored urea denaturation experiments were collected on an AVIV CD spectrometer and on an Applied Photophysics Chirascan instrument. Lyophilized protein was dissolved in 20 mM sodium acetate buffer with 100 mM NaCl with and without 10 M urea. The protein concentration was 18 - 22 µM. Denaturations were performed at 25°C. The concentration of urea was determined by measuring the refractive index. The concentration of protein was estimated using the FCN absorbance measured at 280 nm based on an extinction coefficient (ϵ) of 850 $\mathrm{M}^{-1}\,\mathrm{cm}^{-1}$ (for donor only mutants) or the combined FCN and Trp absorbance measured at 280 nm and based on a total ϵ of 6350 $\mathrm{M}^{-1}\,\mathrm{cm}^{-1}$ (for donor/acceptor mutants). Denaturation curves were fit to equations 7.1 and 7.2

$$\theta_{222} = \frac{a_n + b_n[DEN] + (a_d + b_d[DEN] \exp\left(\frac{-\Delta G^{\circ}([DEN])}{RT}\right)}{1 + \exp\left(\frac{-\Delta G^{\circ}([DEN])}{RT}\right)} \tag{7.1}$$

Where

$$\Delta G^{\circ}([DEN]) = \Delta G^{\circ}(\mathrm{H_2O}) - m[DEN] \tag{7.2}$$

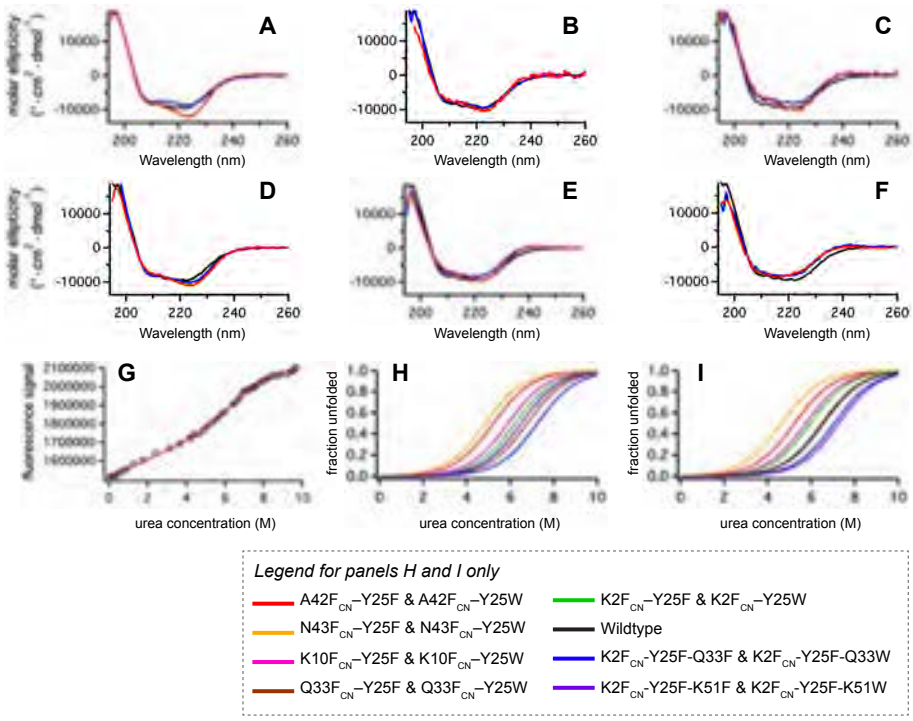

Figure 7.1: (A) to (F) shows CD spectra of FCN-Trp variants compared to wild-type NTL9. Wild-type (black), donor only (blue), donor + acceptor (red). (A) Q33FCN-Y25F (blue), Q33FCN-Y25W (red) (B) K10FCN-Y25F (blue), K10FCN-Y25W (red) (C) A42FCN-Y25F (blue), A42FCN-Y25W (red) (D) N43FCN-Y25F (blue), N43FCN-Y25W (red) (E) K2FCN-Y25F (blue), K2FCN-Y25W (red) (F) K2FCN-Y25F-Q33F, K2FCN-Y25F-Q33W (red). Spectra were recorded at 25 °C in 20 mM sodium acetate and 100 mM NaCl at pH 5.5. The protein concentration was 15-25 µM. (G) Steady-state fluorescence monitored urea denaturation of K10FCN-Y25F. The protein was dissolved in 120 mM sodium acetate buffer at pH 5.5. The protein concentration was 17 µM and the experiment was performed at 20°C. (H) and (I) show plots of the fraction unfolded versus urea concentration. (H) Donor only (I) Donor + acceptor with colors as defined in the legend.

7.2.3 Equilibrium Time-resolved Fluorescence

Time-resolved fluorescence experiments were performed at the University of Massachusetts Medical School in Worcester, MA. All plastic microcentrifuge tubes, conical tubes, pipet tips, and 96-well plates used in sample preparation were sonicated in and washed extensively with 2% Hellmanex III solution (Helma Analytics) and distilled water prior to use in order to reduce background fluorescence arising from impurities on the plastics. For equilibrium experiments, dry protein was dissolved in 120 mM sodium acetate buffer at pH 5.5 with or without 10 M urea for a final protein concentration of 19-25 µM. NaCl was not used since Cl- quenches FCN fluorescence and would complicate data analysis. The folding stability and rate of NTL9 is the same under the modified buffer conditions. The concentration of urea was determined by measuring the refractive index. The concentration of donor only proteins was estimated using the FCN absorbance measured at 280 nm and the concentration of donor/acceptor proteins was estimated using the combined absorbance of FCN and Trp measured at 280 nm. Samples at different urea concentrations were loaded on a 96-well plate using an automated titrator. Fluorescence lifetime measurements were performed on a home-built time correlated single photon counting (TCSPC) apparatus. FCN was preferentially excited at 240 nm using the tripled output of a 10 W Verdi (Coherent) pumped Ti:sapphire laser (Coherent Mira). The repetition rate was 3.8 MHz. Fluorescence was collected through a bandpass filter (FF01-292/27, Semrock, Rochester, NY) and a Glan-Taylor polarizer at the magic angle. A PMH-100-6 photomultiplier tube connected to a SPC150 photon counting card (Becker-Hickl, Berlin, Germany) was used for TCSPC. All measurements were made at 20 ± 1 °C.

7.2.4 Continuous-Flow Time-Resolved Fluorescence

For all kinetic experiments, lyophilized proteins were dissolved in 120 mM sodium acetate buffer with 9.5 M urea at pH 5.5 and passed through a 0.22 µm syringe filter during injection into the sample loop. The refolding buffer was 120 mM sodium acetate at pH 5.5 with no urea. All flow rates to the mixer were adjusted so as to achieve a 10-fold dilution for a final urea concentration of 0.95 M and final protein concentrations between 15 and 25 µM. The initial protein concentrations were 300 µM for K2FCN-Y25F, Q33FCN-Y25F and Q33FCN-Y25W and 150 µM for K2FCN-Y25W in 20 ml of 9.5 M urea.

A 50 µm × 100 µm quartz mixer was used. Flow to the mixer was provided by two syringe pumps (Isco) operating at a combined flow-rate of $4\,\mathrm{ml\,min^{-1}}$ which corresponds to a linear velocity of $82.5\,\mathrm{\mu s\,mm^{-1}}$. Protein solutions in 9.5 M urea were pumped into the mixer by buffer with the same urea concentration at a flow-rate of $0.4\,\mathrm{ml\,min^{-1}}$. The protein solution was mixed with dilution buffer which was pumped at a flow-rate of 3.6 ml/min. The time for complete mixing of solutions was 70 µs. Decays were collected in 7 µs step sizes and all decays up to 1963 µs were used in the analysis. For refolding of K10FCN and N43FCN pairs, protein was dissolved in 20 ml urea buffer to a concentration of 250 µs. A 50 µm × 100 µm quartz mixer was used. The combined flow-rate to the mixer was $4\,\mathrm{ml\,min^{-1}}$ which corresponds to a linear velocity of $82.5\,\mathrm{\mu s\,mm^{-1}}$. Protein solutions in urea were pumped into the mixer by buffer at the same urea concentration at a flow-rate of 0.4 ml/min and the dilution buffer was pumped at a flow-rate of 3.6 ml/min. The time for complete mixing of solutions was 85 µs. Decays were collected in step sizes of 9 µs and all decays up to a folding time of 2404 µs were used in the analysis. For experiments with the N42FCN constructs the initial protein concentration in 10 M urea was 250 µs. A 70 µs × 100 µs quartz mixer was used.

Protein in 9.5 M urea was pumped into the mixer at a flow-rate of 0.6 ml/min and diluting buffer was pumped at 5.4 ml/min for a combined flow-rate of 6 ml/min. This corresponds to a linear velocity of 85 µs mm^{-1}. Time for complete mixing was 41 µs. Decays were collected in step sizes of 41 µs and all decays up to a folding time of 2348 µs were used in the analysis. K2FCN-Y25F-Q33F/W constructs were dissolved in 9.5 M urea for an initial concentration of 500 µM. A 70 µm × 100 µm quartz mixer was used. The protein in 9.5 M urea was pumped through the mixer at 0.3 ml min^{-1} and diluting buffer was pumped at 2.7 ml/min for a combined flow-rate of 3 ml/min. The time for complete mixing was 68 µs. Decays were collected in step sizes of 34 µs and all decays up to a folding time of 2059 µs were used in the analysis.

7.2.5 Equilibrium SAXS

All small angle x-ray scattering experiments were performed at the Advanced Photon Source at the Argonne National Laboratory in Illinois. Scattering profiles were collected for wild-type NTL9 as a function of urea concentration between 0 and 10 M urea. The protein concentration was 3.3 mg ml^{-1} in 120 mM sodium acetate at pH 5.5 and varying concentrations of urea.

7.2.6 Continuous-Flow SAXS

A total of 4 re-folding experiments were performed to obtain adequate signal to noise. Dry, wild-type NTL9 protein was dissolved in 120 mM sodium acetate at pH 5.5 with 8 M urea for a final protein concentration of 25-35 mg ml^{-1}. NTL9 was also dissolved in 1 M urea for a final protein concentration of 4 mg ml^{-1} to collect the scattering curve for the folded state

endpoint. Protein solutions were filtered through a 0.22 µM syringe filter during injection into a 5 ml sample loop. Protein solution in 8 M urea was pumped into the mixer at a flow-rate of 0.5 ml min^{-1} and diluting buffer was pumped at 3.5 ml/min to achieve a final urea concentration of 1 M and final protein concentration of 3.1-4.4 mg ml^{-1}. The time for complete mixing was 178 µs and scattering profiles were collected in 13 µs step sizes. Scattering curves collected between 178 µs and 3976 µs were used in the analysis. Experiments were performed at 25°C. NaCl was eliminated so that samples were collected under the same conditions as in FRET experiments. The concentration of urea was determined by measuring the refractive index. The concentration of protein was estimated using the Tyr absorbance measured at 280 nm based on an ϵ of 1490 /M/cm.

7.2.7 Recording and Analysis of Time Resolved Fluorescence

The folded state florescence lifetime data was fit to a Gaussian distribution, with the mean distances range from 8 Å for the 33FCN-W25 pair to 21 Å for the 42FCN-W25 and 22FCN-W33 pairs. The high denaturant unfolded state was fit separately to a Gaussian distribution and to a wormlike chain (WLC) model, taking into account diffusion during the excited-state lifetime. Both models have been used to fit FRET data, but our results here show no dependence on which of the models is used, providing confidence that the underlying model is not biasing the derived distances [216, 428, 552].

Singular value decomposition (SVD) analysis of the continuous-flow data in 1 M urea gives a maximum of two components along both the folding and time correlated single photon counting (TCSPC) axis. The amplitude of the major and minor SVD components along the folding time axis fit well to a single exponential model, consistent with two-state folding (fig.

7.7). This allowed us to globally analyze the continuous-flow data using a two-population model, folded and unfolded. The observation that the data is well fit using two SVD components does not mean that the unfolded state properties are independent of urea, and clear differences between the 10 M and 1 M unfolded states are observed. As before, the unfolded states were fit to Gaussian distribution and the WLC model with diffusion. The WLC model contains only the amplitude and persistence length as fitting parameters and this allows a diffusion parameter to be introduced without over-parameterizing the fits. Although the absolute value of the distances is slightly different between the Gaussian distribution and WLC models, fits to both models with and without diffusion indicate significant compaction of the protein upon dilution out of high denaturant. The greatest extent of collapse is seen for the two variants that probe the largest sequence separation, 2-25 and 2-33 (fig. 7.9). The mean distance in the unfolded state derived from WLC analysis of the Fcn2-Trp25 pair is 36.5 Å in 10 M urea and 21.4 Å in 1 M urea, while the mean distances observed for the Fcn2-Trp33 pair are 40.1 Å in 10 M urea and 24.1 Å in 1 M urea.

7.2.8 Analysis of Equilibrium Fluorescence Data

To determine the ΔG° and m-values for each construct, SVD analysis was performed on each data set containing fluorescence lifetime decays as a function of urea concentration. All equilibrium denaturations were described by a minimum of one to three components. The amplitudes of the components were fit globally to a two-state folding model where the ΔG° and m-values. The m-value was either a freely floating parameter or fixed to the wild-type value of $0.66\,\mathrm{kcal\,mol^{-1}\,M^{-1}}$ kcal mol-1 M-1.

The instrument response function was obtained by either recording a scattered light signal or by maximum entropy numerical deconvolution from the fluorescence decay of free FCN, which is single exponential. Donor only decays were fit to single or double exponential models re-convoluted with the instrument response function in order to obtain the lifetime values in the absence of acceptor which were subsequently fixed in the analysis of the donor/acceptor decays. Donor/acceptor decays were fit several ways. One method involved an analytic function described by a Gaussian for the native state (see eq. 7.3 and 7.4):

$$I_{da}(t) = \int_0^{l_c} l_d(t)p(r)\exp\left\{-\frac{t}{\tau_D}\left[1+\left(\frac{R_0}{r}\right)^6\right]\right\}dr \tag{7.3}$$

Here l_c is the contour length between the donor and acceptor in Å and is calculated as $3.8N$, where the prefactor of 3.8 Å corresponds to the distance between two consecutive Cα atoms and N is the number of amino acids between the donor and acceptor. τ_D is the lifetime of the donor in the absence of acceptor, R_0 is the Förster radius which was previously determined to be 16 Å for the FCN-Trp pair5, r is the distance in Å and the distance distribution $p(r)$ is

$$p(r) = a\exp\left(\frac{-(r-\omega)^2}{2\sigma^2}\right) \tag{7.4}$$

where $p(r)$ is the probability of finding the donor and acceptor pair separated by r, a is the amplitude, ω is the center of the distribution in Å, and σ is the width of the distribution in Å. The adjustable parameters in the fits are a, ω and σ. Donor/acceptor decays were fit

using either a Gaussian (eq. 7.3) or the Ha-Thirumalai wormlike chain (WLC) model for the unfolded state (eq. 7.5)

$$p(r) = \frac{4\pi N r^2}{l_c^2 \left(1 - \left(\frac{r}{l_c}\right)^2\right)^{9/2}} \exp\left(\frac{-3l_c}{4l_p[1 - \left(\frac{r}{l_c}\right)^2]}\right) \tag{7.5}$$

where N is the normalization constant, l_c is the contour length in Å, and l_P is the persistence length in Å. The adjustable parameters in the fits are the l_p and the amplitude of the distribution. Diffusion between donor and acceptor during the excited-state lifetime of the donor was modelled using:

$$\frac{\partial \bar{N}(r, t_{TCSPC})}{\partial t} = \left\{\sum_i \frac{a_i}{\tau_i}\left[1 + \left(\frac{R_0}{r}\right)^6\right]\right\} \cdot \bar{N}(r, t_{TCSPC}) + \frac{1}{N_0(r)}\frac{\partial}{\partial r}\left[N_0(r) D(r) \frac{\partial \bar{N}(r, t_{TCSPC})}{\partial r}\right] \tag{7.6}$$

where

$$\bar{N}(r, t_{TCSPC}) = \frac{N^*(r, t_{TCSPC})}{N_0(r)} \tag{7.7}$$

In this equation, N^* is the excited state population at time t_{TCSPC} following excitation, N_0 is the distribution at $t = 0$ and D is the diffusion coefficient.

7.2.9 Analysis of Continuous-Flow Time-Resolved Fluorescence Data

Decays for the donor only protein and for the donor/acceptor were each initially processed using the CF-TCSPC data processor written in LabVIEW (National Instruments). Summations, corrections, subtractions and SVD analysis were performed using this software. The amplitudes of the major and minor SVD components as a function of folding time fit well to a single-exponential model indicating that folding is two-state, and that the data can be analyzed in terms of two populations, folded and unfolded. The data matrix was then exported and loaded into the software package Savuka where global analysis was performed on the decays [209]. Donor only decays as a function of folding time were fit globally to obtain the lifetimes in the absence of acceptor. Lifetimes across all decays were linked and remained constant as a function of folding time, but the fractional amplitudes of each lifetime component varied. Donor only lifetimes and fractional amplitudes were fixed during analysis of donor/acceptor data. Donor/acceptor fluorescence decays were fit using two distributions. The distance distribution for the folded state was modeled using a Gaussian distribution while the distance distribution for the unfolded state was modelled using either a Gaussian distribution or WLC model. Diffusion in the unfolded state was taken into account when fitting using the WLC model. Rapid compaction of the unfolded state occurs within the dead time of the instrument and thus the high urea unfolded state is not observed. Parameters of the folded and unfolded distance distributions (mean, FWHM, l_p, and the diffusion coefficient) were linked and remained constant as a function of folding time, while the relative amplitudes of the folded and unfolded distributions (populations of each state) varied.

7.2.10 Analysis of Equilibrium SAXS Data

In order to obtain the radius of gyration (R_G), the Kratky plot, and the pairwise distance distribution function $P(r)$ at high urea, scattering curves collected at 9, 9.2, 9.3, 9.5, 9.6, 9.8 and 9.9 M urea were averaged. Scattering curves contained 1308 points along the q-axis and every three points were averaged so that the final scattering curve contained 433 data points. The $P(r)$ was determined using the data analysis software ATSAS8. The radius of gyration was obtained using the Guinier approximation:

$$I(q) = I(0) \exp\left(\frac{-R_G^2 q^3}{3}\right) \tag{7.8}$$

Where $I(q)$ is the intensity at scattering vector q. The Guinier approximation is valid for qq_{max} with $q_{max} \times R_G \sim 1.3$ for globular molecules such as folded proteins, and $q_max \times R_G \sim 0.8$ for extended molecules. The range is thus more limited for unfolded proteins than for folded proteins. Using $q_{max} \times R_G$ between 0.90 and 1.08 (10 fewer/greater number of data points) provided a similar value for R_G (23.4 - 23.7 Å). The R_G becomes progressively smaller if $q_{max}\ R_G > 1.08$ is used, and is 22.6 Å if the range is extended to that commonly used for globular particles, $q_{max} \times R_G = 1.26$.

7.2.11 Analysis of Continuous-flow SAXS Data

Four refolding experiments were performed on wild-type NTL9 in which scattering curves were measured as a function of refolding time. Scattering curves were collected between 50.8 and 3976 µs in step sizes of 12.7 µs. The earliest time-point included in the analysis was 177.8 µs. The dataset was fit globally to a combined kinetic-equilibrium model in order to

determine the extrapolated scattering profiles at $t = 0$ (unfolded state) and $t = \infty$ (folded state), and also to take into account any initial folded state population.

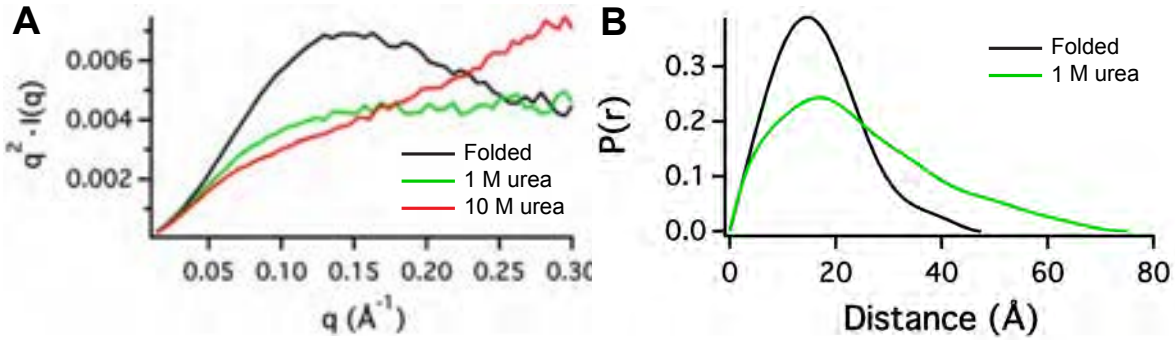

Figure 7.2: (A) Kratky plots for wild-type NTL9. Folded state in 1 M urea (black), unfolded state in 1 M urea (green), unfolded state in 10 M urea (red). Plots for the folded and unfolded proteins in 1 M urea (black and green) are from the global analysis of the continuous-flow refolding experiments. Results for the protein in 10 M urea (red) are from a separate equilibrium experiment. (B) $P(r)$ distribution obtained using the scattering curves determined from the global analysis of the continuous-flow data. Folded state (black) and unfolded state (green) in 1 M urea.

The radius of gyration was obtained using the extrapolated scattering profiles for the folded and unfolded states and the Guinier approximation (equation 7.8). The values obtained for R_G are similar (18.9 - 19.2 Å) if q_{max} R_G is varied from 0.97 to 1.25 (10 fewer/greater number of data points). The pair distribution for the folded and unfolded states in 1 M urea was calculated using ATSAS.

7.2.12 Monte Carlo Simulations

The simulations were conducted using the CAMPARI Monte Carlo simulation engine and the ABSINTH implicit solvent model. Parameters were taken from the abs3.2_opls.prm parameter file. Ten independent simulations were performed at each of the following temperatures: [240, 260, 280, 290, 300, 310, 320, 330, 340, 345, 350, 355, 360, 365, 370, 375, 380, 390, 400, 430, 450, 500]. The ensembles generated at each temperature consisted of 60,000 configurations. Conformations that contained less than 50% native contacts were kept as part of the unfolded ensemble. The simulations were analysed using the CTraj analysis framework (see chapter 9).

Simulations of the excluded volume (EV) ensemble, Flory Random Coil (FRC) ensemble, and Lennard-Jones (LJ) ensemble were performed as described previously (see chapter 5). The EV, FRC, and LJ ensembles generated correspond to an NTL9 specific ensemble behaving as a polymer in a good, Θ, or poor solvent, respectively. For ensembles in the EV limit, all attractive Hamiltonian components are set to zero, while the repulsive components from the Lennard-Jones potential drive chain expansion due to steric repulsion. For ensembles in the FRC limit, the dihedral angles sampled in the FRC simulations were taken from a residue-specific database of known allowed dihedrals. The residue local steric behavior was included and all other Hamiltonian components were set to zero. For ensembles in LJ limit, attractive and repulsive Lennard-Jones Hamiltonian terms are allowed, but all other terms (e.g. solvation, electrostatics, *etc.*) are turned off, such that the only attractive interactions are chain-chain.

7.2.13 Evaluation of Reweighted Ensembles

The ensemble generated at 375 K, upon minor re-weighting, best matched both the FRET and SAXS experimental data for the unfolded state in 1 M urea. Previous results have shown that the simulated ensemble at 390 K matches the 8 M urea unfolded ensemble, and describes how the un-reweighted ensembles were identified [377].

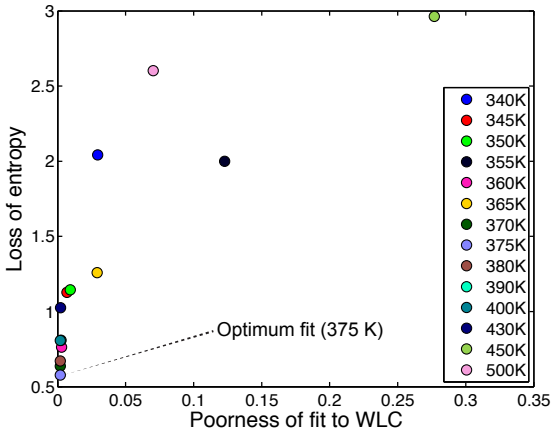

Figure 7.3: A comparison of the loss of entropy and poorness of fit associated with the various reweighted ensembles. The ensemble at 375 K minimizes both the poorness of fit and the loss of entropy. Not all simulation temperature are shown as they are substantially lower in terms of fit quality and loss of entropy.

7.2.14 Procedure for Ensemble Reweighting

The entropy change and the goodness of fit of the simulated histograms to the experimentally determined WLC models varied for the different ensembles. The best fit and smallest decrease in entropy was obtained using the ensemble at 375 K. The loss of entropy is a measure of overfitting and the 10% reduction in entropy ($\Delta S = -0.5$) observed indicates a

small perturbation to the original ensemble. This procedure yields a unique global solution in which the reduced χ^2 value is minimized and the entropy is maximized. This method was developed by Leung *et al.* explicitly for the re-weighting of large ensembles to match experimental data and provides a robust, powerful, and efficient approach [324].

The loss of entropy experienced upon re-weighting was calculated using

$$\Delta S = \left(-\sum p_i^0 \log(p_i^0)\right) - \left(-\sum p_i^{\mathrm{RW}} \log(p_i^{\mathrm{RW}})\right) \tag{7.9}$$

Where i corresponds to each conformation, p_i^0 is the weight for conformation i in the un-weighted ensemble ($1/N$, where N = number of conformations), while p_i^{RW} represents the probability of the ith conformation in the re-weighted ensemble.

The re-weighted ensembles were obtained by re-weighting five of the six distances to match the experimental results. The pairs 25-42 and 25-43 are so close in that we found no difference if both were included, so to minimize the perturbation to the ensemble only one of the two (25-42) was included. The distances at 375 K closely resembled the experimental FRET distances even before reweighting. Four of the six pairs (25-33, 10-25, 25-42, 24-43) are almost indistinguishable from the FRET derived distances (fig. 7.4A - D). The remaining two (2-25 and 2-33) showed a bimodal distance distribution before reweighting (fig. 7.4E and F). One of the peaks overlap with the experimentally determined WLC distribution, while the other was centered at a significantly higher distance. After re-weighting the distribution centered at higher distances was lost and the distribution that overlaps with the experimentally derived WLC model increased in population. The WLC model makes numerous simplifying assumptions and will not be a perfect representation of the distance distributions within the

1 M urea unfolded ensemble due to the structural heterogeneity and the presence of local and long-range contacts, which are not considered in the model.

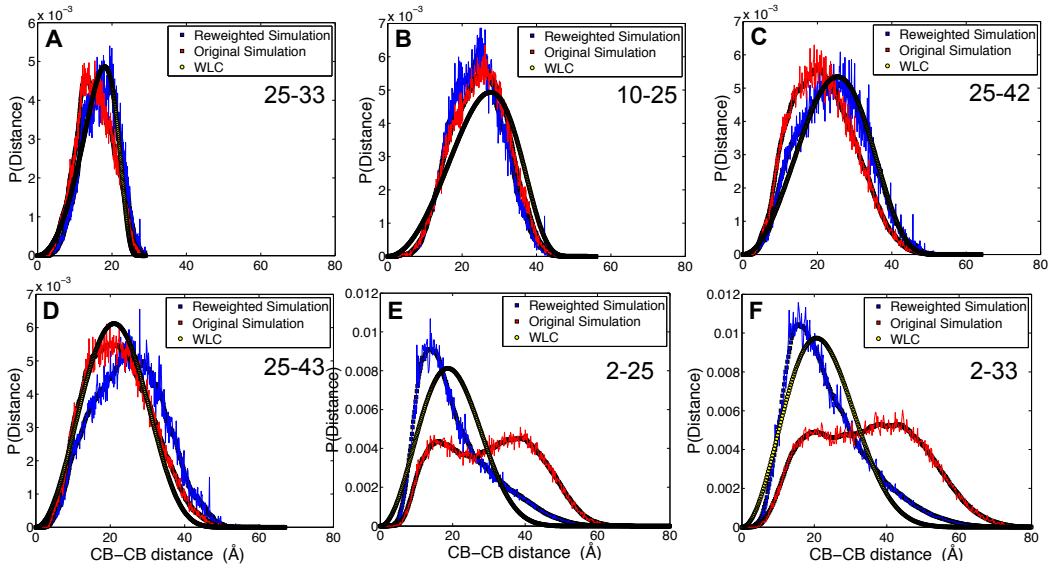

Figure 7.4: The impact of re-weighting shown for each of the distances, comparing the distance distribution between the WLC distribution (yellow), un-weighted simulation (red), and re-weighted simulation (blue). For pairs (A) 25-33, (B) 10-25, (C) 25-42 and (D) 25-43 the un-weighted and re-weighted simulations are extremely close, suggesting these distances do not require significant re-weighting. For the two furthest pairs (E) 2-25 and (D) 2-33, a bimodal distribution emerges from the simulation ensemble. The re-weighting serves to redistribute these two populations, generating a unimodal distribution that is consistent with the WLC model.

7.2.15 Polymer Scaling Analysis in Finite Chains

We used a model free method to determine the scaling exponent ν^{app} (see chapter 2 for an discussion on ν vs ν^{app}. As a reminder, the global scaling behaviour of a finite-length polymer can be written as

$$R_G = A_0 N^{\nu_{app}} \tag{7.10}$$

For heteropolymers, we can analyze the scaling behaviour using internal scaling profiles. This approach determines ν^{app} and A_0 using the following relationship[24]

$$\log(\langle\langle r_{i,j}\rangle\rangle = \nu^{app} \log(|i-j|) + A_0 \tag{7.11}$$

Using this relationship and the set of data for r_{ij} and $|i-j|$ generated by the ensembles, we scan across all reasonable combinations of A_0 and ν^{app} to identify the pairwise combination of A_0 and ν^{app} to that gives rise to the best fit to the data. The resulting fitting-landscape is convex, with a single minimum corresponding the best ν^{app} and A_0 values.

Polymer scaling theory was developed in the context of polymers of infinite length. While it has proven to be remarkably powerful for large macromolecules, the applicability of these analyses to *much* shorter polypeptides raises some questions regarding the impact of finite size effects. To determine A_0 and ν^{app}, we found that if all $|i-j|$ sequence separations were

[24]Of note - we find essential identical results when using $\sqrt{\langle\langle r_{i,j}^2\rangle\rangle}$ instead of $\langle\langle r_{i,j}\rangle\rangle$, and have extensively explored the various options for fitting to internal scaling behaviour in many different systems. The $\sqrt{\langle\langle r_{i,j}^2\rangle\rangle}$ is formally correct, and will be used for scaling exponent estimation going forward

used, the derived scaling exponents for well-defined ensembles (e.g. polymers in a good or Θ-solvent) were slightly too large, a result consistent with previous work, and due to the fact that a significant number of the $|i-j|$ distances are too short to experience true scaling behaviour [377]. In addition to this, we found a strong dependence on the 'dangling ends' of the chain, due to the fact that there is only a single pair of residues that shows $|i-j| = N_{res}$, such that if either of the end-residues engage in specific intramolecular interaction this can provide an unreasonably highly weighted disruption to the apparent scaling behaviour.

To account for this, we identified two corrections to be performed when fitting the internal scaling data. Firstly, we identified a minimum threshold $|i-j|$ where a good-solvent simulation gave a scaling exponent of 0.59 ($|i-j|$ threshold = 15), and verified that this minimum threshold could reproduce scaling exponent of ∼0.5 for a Θ-solvent simulation. We also discarded the five largest sequence separations (i.e. $|i-j| > (N_{res}-5)$ to avoid the dangling-ends having a significant impact on scaling behaviour, although. We then analysed all simulation data in a self-consistent manner. Consequently, while we have identified clear trends in A_0 and ν^{app}, the exact values should be treated as qualitative estimates, as opposed to concrete and exact values.

The scaling exponent ν^{app} identified in this approach is identical to ν in equation 7.10. However, A_0 is, qualitatively, a factor of $\sim\sqrt{6}$ different from R_0. Consequently, for the 1 M urea DSE R_0 is approximately 2.6 (slightly larger than the values reported for fully unfolded proteins in this study and in previous studies [297]). Setting A_0 to 2.6 and ν to 0.48 and using the relationship defined in equation 7.10 yields an expected R_G of 18.0 Å, in good agreement with the value obtained from SAXS (R_G = 19.1 Å) and simulation (R_G 18.9 Å). We hypothesize that this slightly larger A_0 value reflects an increase in effective monomer

size brought about by the formation of local structure, which also contributes to an effective increase in persistence length.

7.3 Results

NTL9 is 56 residues in length and is one of the simplest examples of a common mixed α-β fold, known as the split $\beta-\alpha-\beta$ motif [233]. The domain folds in a two-state fashion under a broad range of conditions and has been shown to contain residual structure in the unfolded state populated under native conditions [10, 98, 308]. The folding time constant of wild-type NTL9 in 1 M urea, pH 5.5 is 2.4 ms, allowing much of the folding process to be accessed using continuous-flow methods. NMR experiments have been conducted using a variant with a pair of destabilizing mutations in the hydrophobic core, resulting in an unfolded population of 70% under native conditions that is in slow exchange with the folded state. Analysis of the NMR data showed that the unfolded state under native conditions contains residual α-helical secondary structure and both native and non-native long-range contacts between hydrophobic residues [376].

7.3.1 FRET Constructs Show Wildtype-Like Stability & Folding Rates

A set of seven variants of NTL9 containing FCN and Trp were prepared using 21st pair technology. Sites were selected to probe a range of positions and to ensure that FRET pairs span a range of sequence separations (fig. 7.5).

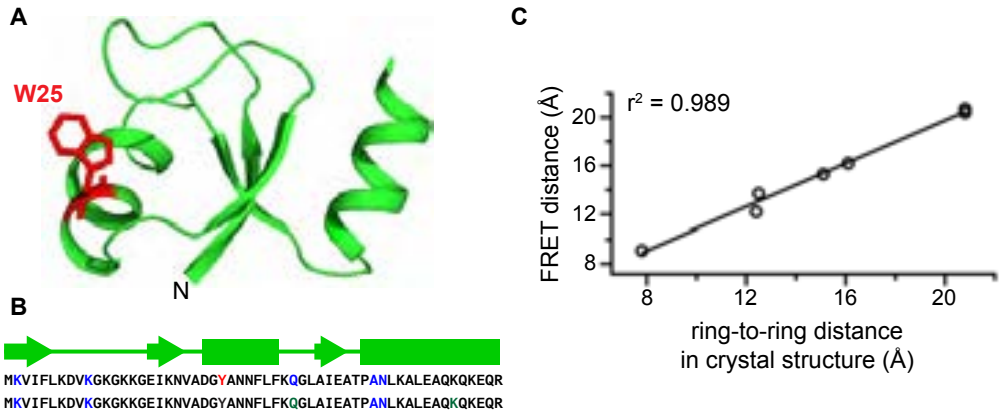

Figure 7.5: (A) Ribbon diagram of NTL9 native state (PDB code: 2HBB) showing the location of Trp-25. The N terminus is labelled. (B) Primary sequence of NTL9. (Top) The location of Trp-25 is shown in red and the location of the 5 FCN substitutions which are paired with Trp-25 are colored blue. (Bottom) Residues 2 and 33 as well as residues 2 and 51 were used to prepare additional FCN-Trp pairs and the position of these Trp residues is colored green. The schematic diagram below the sequence illustrates the secondary structure. (C) The FRET derived distances for the native state of NTL9 compared to the ring-to-ring distances taken from the crystal structure.

The sequence separation between donor and acceptor varies from 8 to 49 residues, while the distance between the sites (Cβ-Cβ) varies from 11 to 21 Å in the native state. All seven were well folded, displayed sigmoidal unfolding profiles, and showed similar CD profiles to the wildtype protein (fig. 7.1). The distances derived from equilibrium FRET measurements of the folded protein under native conditions show perfect agreement with the expected distances derived from the crystal structure (fig. 7.5), demonstrating that the dyes do not perturb the folded tertiary structure, and that we are able to extract meaningful distances from the FRET distributions. Taken together these data suggest that the dyes are non-perturbing, yet offer precise local distance information.

7.3.2 The Unfolded State in 10 M Urea is Expanded

We first characterized the folded state and the 10 M urea unfolded state at equilibrium to establish baselines for comparison to the data collected in continuous-flow mode. Folded state measurements were made in 1 M urea since this corresponds to the final conditions in the continuous-flow mixing studies, and represents a 'native' condition where the protein is ∼100% folded (fig. 7.1). The fluorescence lifetime data was fit to extract distance distributions between donor and acceptor in the folded state and in the high urea unfolded state (fig. 7.6). The folded state was fit to a Gaussian distribution and gives, as expected, narrow distributions (fig. 7.9) that show excellent agreement with the crystal structure (fig. 7.5C). The high denaturant state was fit separately to both a Gaussian distribution and wormlike chain (WLC) model with diffusion, with both fitting methods yielding nearly identical results. The distributions for the unfolded protein are much broader and show a monotonic increase in mean distance vs. sequence separation, as expected for an expanded chain in high denaturant (fig. 7.8).

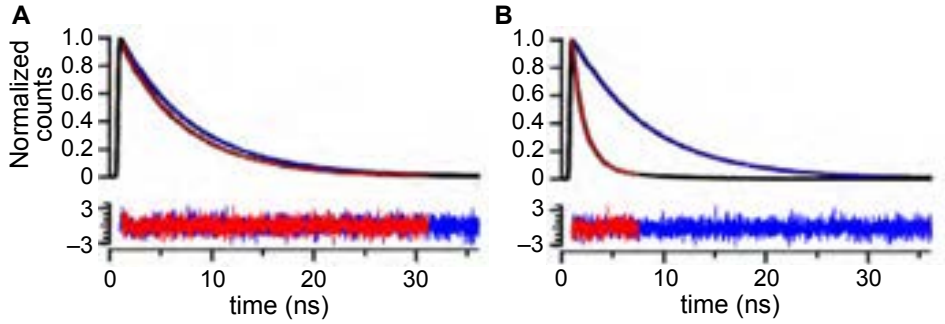

Figure 7.6: Representative fluorescence lifetime measurements. Data is shown for the FCN10-W25 pair. Donor only curves are in blue, donor plus acceptor in red. Residuals are shown below. (A) Equilibrium data in 0 M urea. (B) Equilibrium data in 10 M urea.

7.3.3 The Unfolded State Populated in Low Concentrations of Urea is More Compact

Fluorescence lifetime measurements were conducted in combination with continuous-flow mixing to allow fluorescence decays to be collected as a function of refolding time (fig. 7.7). Rapid collapse is observed within the 50 - 100 µs dead time of the instrument. The wild-type NTL9 folding time in 1 M urea is 2.4 ms, indicating that collapse occurs on a much faster time-scale than folding. Singular value decomposition (SVD) analysis yields a maximum of two components, the amplitudes of which fit well to a single exponential model, consistent with two-state folding, allowing a global analysis using a two-state model.

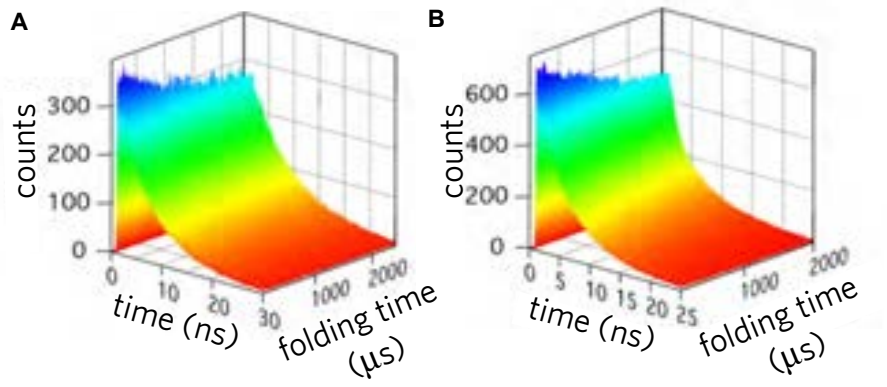

Figure 7.7: Continuous-flow fluorescence decays as a function of folding time for the FCN10, W25 pair. The dead time was 85 µs. Decays were fit globally to a two population model. A Gaussian distribution was used for the folded state and a wormlike chain model for the unfolded state.

The unfolded state contracts rapidly and fluorescence data collected at the earliest time points is already sampling a more compact unfolded state than in 10 M urea. Data for each variant was fit globally using a Gaussian distribution for the folded state and a WLC model or a Gaussian distribution for the unfolded state. Both models suggest significant compaction of the unfolded state upon dilution to 1 M urea. Most notably, the two pairs at greatest sequence separation (FCN2-Trp25 and FCN2-Trp33) exhibit a 40% contraction in the unfolded state in 1 M urea relative to the 10 M urea unfolded state. The changes observed for the other FRET pairs range from 12 to 31%.

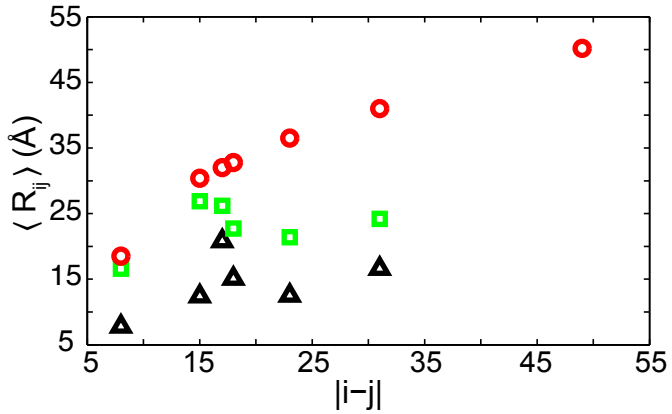

Figure 7.8: The unfolded state is expanded in high denaturant but is more compact in low denaturant. Mean distance versus the separation in primary sequence. NTL9 unfolded in 10 M urea (red circles), NTL9 unfolded state populated in 1 M urea (green squares), and NTL9 folded in 1 M urea (black diamonds).

Figure 7.8 describes the relationship between sequence separation and spatial separation for the protein in 1 M and 10 M urea. At 10 M urea the profile generated is largely consistent with that of a polymer in a good solvent. At 1 M urea, the non-monotonic increase in distance with sequence separation is a hallmark of well-defined anisotropic interactions indicative of a more compact conformation, and is inconsistent with a homopolymer in a good solvent or

Θ-solvent. In summary, the FRET results suggest that significant and rapid collapse occurs upon dilution from 10 M to 1 M urea.

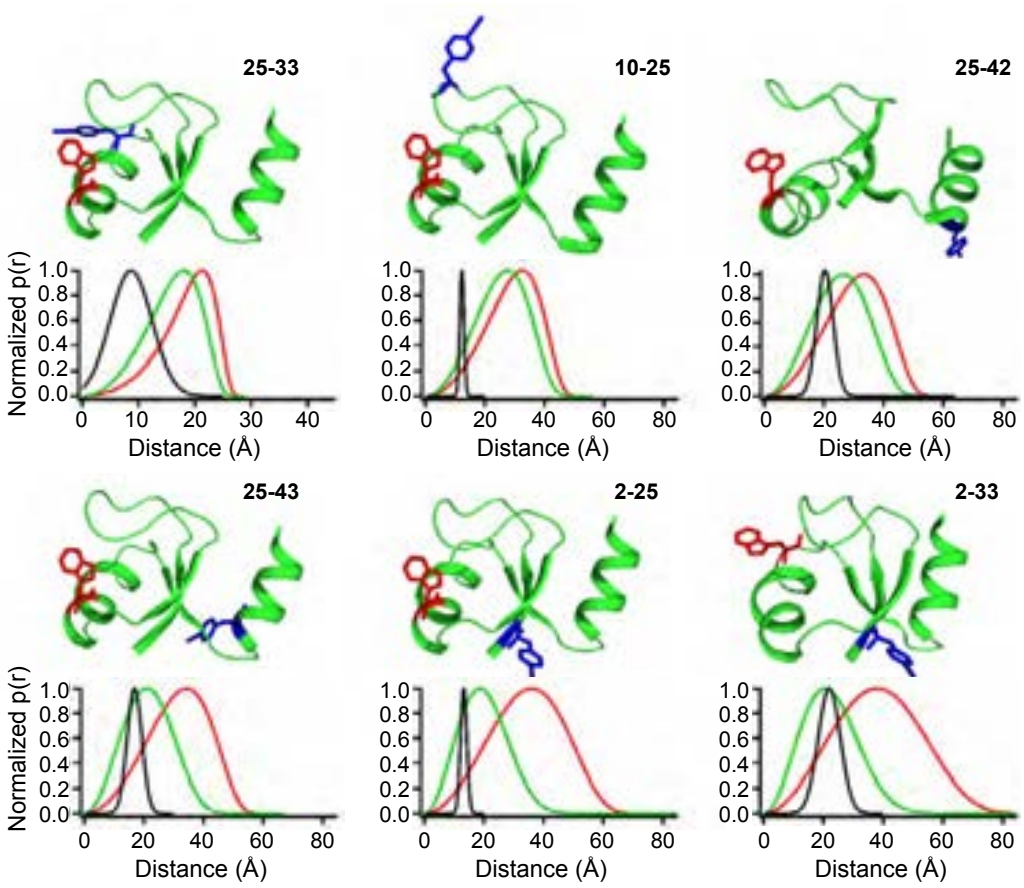

Figure 7.9: FRET provides evidence for compaction. Ribbon diagrams illustrating the location of the FRET pairs are shown together with the distance distributions. Red: Unfolded state in 10 M urea, Green: Unfolded state in 1 M urea, Black: Folded state in 1 M urea. The folded and unfolded distributions in 1 M urea were extracted from the global fit to the FRET data. The unfolded state was modeled as a wormlike chain and a Gaussian distribution was used to model the folded state distribution.

Continuous-flow SAXS data was also collected on wild-type NTL9. A major technical challenge with these experiments originate from the combination of the small dimensions of protein and the presence of urea, meaning the contribution of the protein to the total scattering is low. Protein, initially in 8 M urea, was diluted 8-fold to a final concentration of 1 M urea. The R_G of NTL9 in 10 M urea was determined from equilibrium experiments to be 23.5 ± 0.7 Å which is in excellent agreement with prior equilibrium SAXS studies of the urea unfolded state [376, 377]. The measured R_G for the folded state in 1 M urea is 12.8 ± 0.2 Å and for the unfolded state in 1 M urea is 19.1 ± 0.9 Å (fig. 7.10). The SAXS data thus indicates a more modest ($\sim$19%) compaction upon dilution out of high denaturant than the FRET data does. The Kratky plot for the folded state shows a peak, characteristic of a globular conformation (fig. 7.2). Conversely, a monotonic increase with increasing scattering angle is seen for the unfolded state in 10 M urea, indicating a highly-expanded chain (fig. 7.2). Similar behaviour has been observed for other globular proteins in high concentrations of urea and for IDPs with a high proline content [63, 202]. The Kratky plot for the unfolded protein in 1 M urea is very different, and a plateau is observed at increasing q values indicating a less extended ensemble. The SAXS result suggest that the unfolded state under native conditions is somewhat more compact than at 10 M urea, but is still relatively expanded.

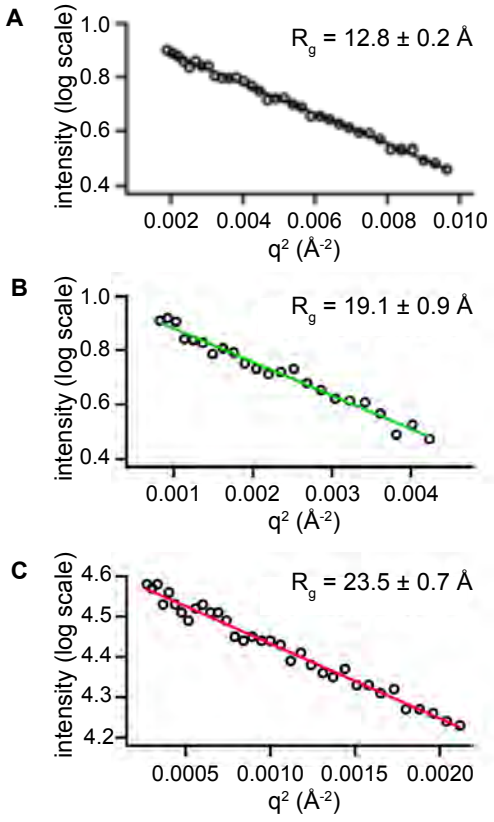

Figure 7.10: Guinier analysis of SAXS data. (A) Continuous-flow data for the native state in 1 M urea. (B) Continuous-flow data for the unfolded state in 1 M urea. (C) Equilibrium data for the unfolded state in 10 M urea.

7.3.4 Simulations Demonstrate that SAXS & FRET are Consistent

The time resolved FRET and SAXS results for NTL9 in 1 M urea suggest seemingly contradictory results. The FRET data suggest a substantial collapse coupled with the formation of well-defined interactions, causing a deviation in intra-chain distances from the expected behaviour for a polymer in either a Θ-solvent or a good solvent. In contrast, SAXS results demonstrate a modest contraction, but are inconsistent with a sharp collapse. To explore this apparent discrepancy, we used all-atom Monte Carlo simulations with the ABSINTH implicit solvent model to generate a series of denatured state ensembles (DSEs) and asses if the results from FRET and SAXS are mutually compatible.

We generated denatured state ensembles of NTL9 at a range of temperatures. We had previously used this approach to show that the DSE generated at 390 K is as a good proxy for the solution behaviour of NTL9 in 8 M urea [377]. Considering this, we sought to identify an ensemble where the derived C-C pair-distances distributions taken from the set of donor/acceptor pairs used for FRET matched the experimentally obtained distance distributions. While several ensembles yielded pair distance distributions that were qualitatively similar to the FRET distance distributions, none were quantitatively identical. To alleviate this issue we used a χ^2 minimization and entropy maximization approach (COPER) to re-weight each denatured state ensemble to match the FRET data [324]. We then selected the re-weighted ensemble that showed the smallest decrease in entropy and best agreement with the FRET data. The best fitting and least perturbed re-weighted ensemble was generated by re-weighting the DSE generated at 375 K, which after re-weighting so good agreement with the FRET derived distances (fig. 7.11 and fig. 7.4). The mean R_G obtained from this re-weighted ensemble (referred to hereafter as the 1 M urea DSE) is 18.9 Å, in excellent

agreement with the value derived from SAXS (19.1 Å, see fig. 7.11B), demonstrating the FRET and SAXS results, while apparently contradictory, and entirely mutually consistent.

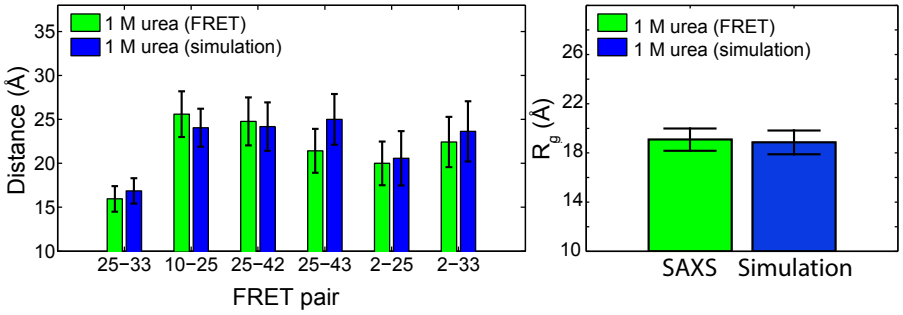

Figure 7.11: Comparison between the experimental results for the 1 M urea unfolded state and the 375 K reweighted ensemble (1 M urea DSE). (A) Intrachain distances obtained from FRET efficiency results (experiment) compared to the CB-CB distance extracted from the unfolded ensemble. (B) Radius of gyration obtained from SAXS (experiment) compared to the ensemble average radius of gyration obtained from simulation. In both cases the simulation-derived values show good agreement with the experimental results, demonstrating that FRET, SAXS, and simulation are mutually compatible.

7.3.5 The DSE in 1 M Urea Experiences Native & Non-Native Interactions

The 1 M urea DSE correctly reproduces the FRET and SAXS results, suggesting it represents a good model for the 1 M urea unfolded ensemble. Considering this, we examined several other structural and polymeric properties of the 1 M urea DSE to gain additional insight into the conformational behaviour of the unfolded state under folding conditions. The ensemble displays native and non-native contacts, extensive but transient long and short-range interactions, and residual native secondary structure (fig. 7.12 and 7.13). These structural

elements are more persistent than is observed for proteins in 8 M urea, and this unfolded state is more compact than is observed for most proteins in a high concentration of denaturant [20]. Taken together, these results suggest that, far from a random coil, the unfolded state under native conditions shows sequence-specific native and non-native conformational and structural preferences.

7.3.6 The DSE Shows Θ Solvent-Like Behaviour Under Folding Conditions

The scaling exponent ν quantifies the relationship between the global dimensions of a polymer and the degree of polymerization (number of monomers). This relationship is described by,

$$\sqrt{\langle R_G^2 \rangle} = A_0 N^{\nu_{app}} \tag{7.12}$$

where R_G is the radius of gyration, A_0 is a prefactor, N is the number of monomers in the chain (amino acids in a protein). In a poor solvent, ν is approximately 1/3 with chain-chain interactions being preferred over chain-solvent interactions, leading to a compact and globular ensemble. In a theta (Θ) solvent, polymer-polymer and polymer-solvent interactions are perfectly counterbalanced and ν is approximately 1/2, leading to large conformational fluctuations and a maximally heterogeneous ensemble [348]. A chain in the Θ-limit behaves as a Flory Random Coil (FRC), also known as an ideal chain or a random flight chain [360]. In a good solvent, ν is $\sim$ 3/5 and the chain is highly expanded due to favourable chain-solvent interactions. A protein in a high concentration of denaturant behaves globally as

a chain in a good solvent, while folded proteins can be described by a scaling relationship consistent with a chain in a poor solvent [144, 297, 377].

We generated an NTL9 specific Θ-solvent ensemble as described previously (see chapter 5 for further discussion). The 1 M urea DSE shows ensemble-averaged internal scaling behaviour consistent with a polymer in a Θ-solvent (fig. 7.12D). To further explore this result, we determined that $\nu \approx 0.48$ for the 1 M urea DSE. Local and non-local intramolecular interactions cancel out across the ensemble-averaged chain to yield global properties consistent with a chain in an approximate Θ-solvent. This would result in experimentally determined global ensemble average properties such as the R_G and the Kratky plot from SAXS to yield results similar to those expected for an ideal chain of equivalent length. The FRET-derived distances provide a probe of specific local conformational behaviour, and demonstrate conformational behaviour inconsistent with those expected for a true ideal chain. Taken together, these results suggest that for NTL9 the unfolded state shows global conformational behaviour consistent with a polymer in a Θ-solvent while simultaneously experiencing long and short range native and non-native interactions.

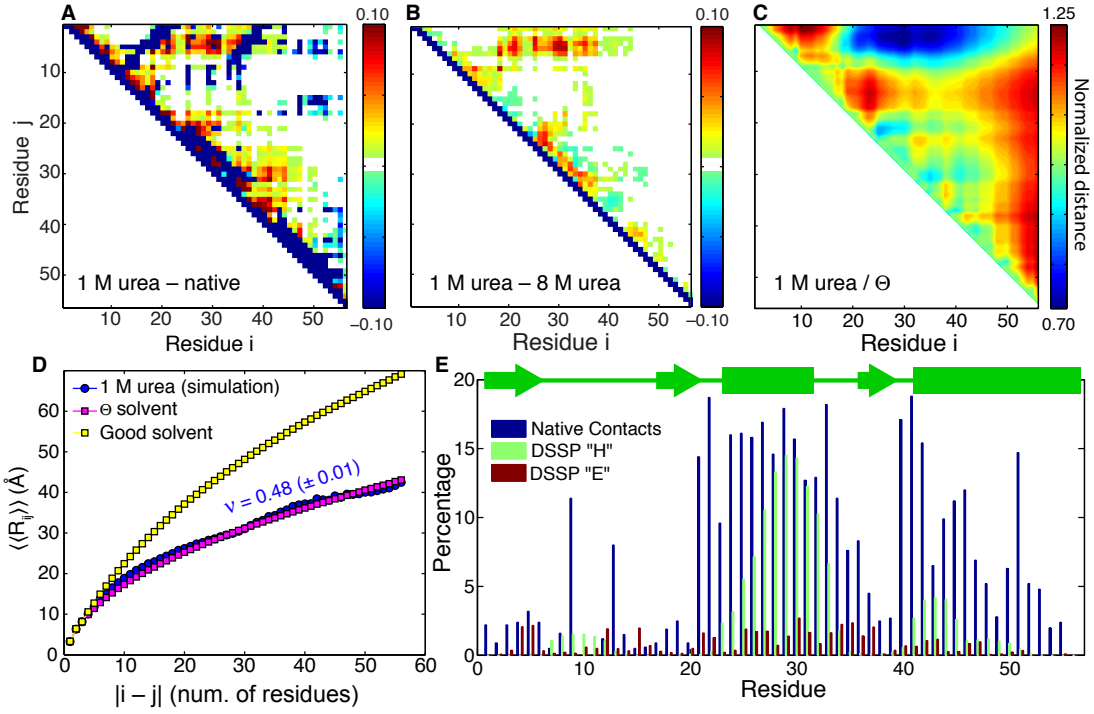

Figure 7.12: Simulation summary figure. (A) Difference contact map (1 M urea - native state contact map). Positive values correspond to non-native interactions. (B) Difference contact map (1 M urea - 8 M urea contact map). Positive values correspond to interactions observed at 1 M but not 8 M urea. (C) Scaling maps normalized by the Θ ensemble scaling map. Cooler colors correspond to regions that are closer together than would be expected in the Θ-state, while hotter colors correspond to regions that are further away. (D) comparison of internal scaling profiles for NTL9 in a good solvent (yellow), Θ-solvent (pink) and from the 1 M urea ensemble (blue). In a globally averaged analysis, the 1 M urea ensemble shows conformational behaviour consistent with a polymer in a Θ-solvent, despite the presence of well-defined secondary and tertiary structure. (E) Presence of native contacts and secondary structure in the 1 M urea ensemble. The native state secondary structure map is shown above in green for reference.

7.3.7 Denatured State Ensembles Show Complex Behaviour

We quantified native contacts, the apparent scaling exponent (ν^{app}), and global dimensions (R_G) for the DSEs generated at 500 K to 280 K (fig. 7.13). Even at 355 K to 375 K, where the un-weighted ensembles show $\nu \approx 0.59$, we observe modest but significant global contraction and the formation of native contacts. While we have no experimental results with which to compare these this analysis to, it is important to highlight that there is no fundamental incompatibility between 'good solvent' scaling behaviour, reduced global dimensions, and the formation of native (and presumably non-native) contacts. Given the extended formation of native structure at these lower temperature (compare 355 K to 375 K) one might expect FRET and SAXS results to becoming increasingly divergent as well defined local conformational preferences cause deviations from homopolymer based statistical models while global, ensemble average behaviours are much less perturbed.

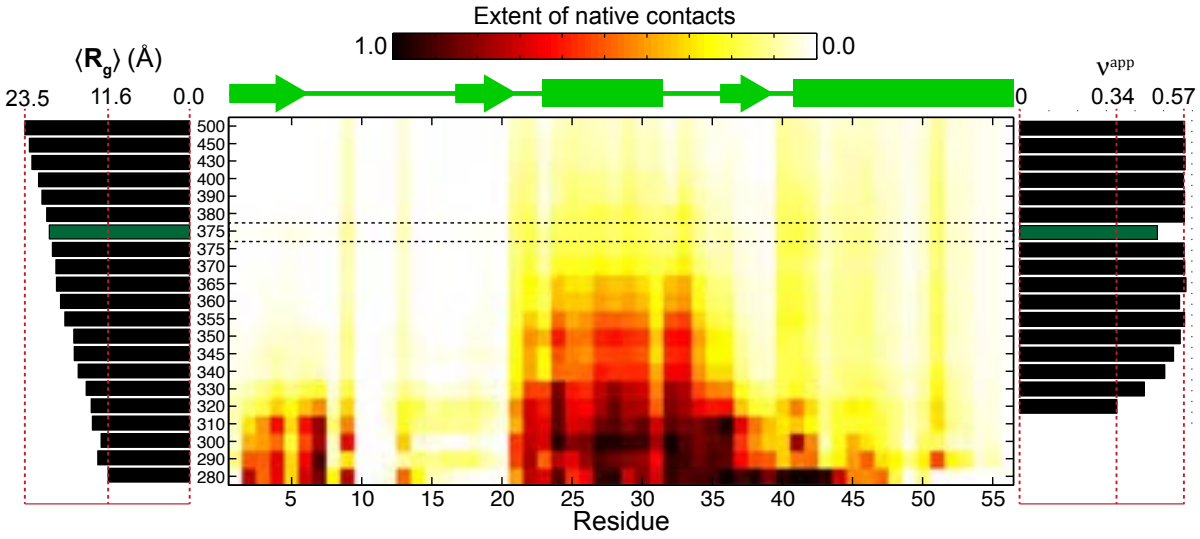

Figure 7.13: For denatured state ensembles generated at 500 K to 280 K we quantify the residue-specific density of native contacts (central heat map), radius of gyration (left hand section) and apparent scaling exponent ν^{app} (right hand section). The reweighted ensemble is highlighted in the dashed box and with green bars. While the 1 M urea ensemble shows native contacts global behaviour consistent with a polymer in a Θ-solvent, ensembles with ν closer to 0.59 (good solvent) are also demonstrate significant native contacts, suggesting that the apparent scaling exponent may not be sensitive to the formation of local contacts and long-range interactions, or a contraction of the radius of gyration, a result consistent with previous work [376].

A_0, an extracted parameter that is directly proportional to R_0 (eq. 7.12) shows a non-monotonic behaviour, with an inflection point at around 350 K (fig. 7.14). The physical origin of A_0 is a convolution of chain persistence length and the volume occupied by a single monomer, suggesting that this may provide a useful order parameter with which to identify the formation of more persistent structure. It also suggests that the often-stated assumption that A_0 (or as it is sometimes written R_0) is solvent-quality independent may not be valid. Finally, this inflection point coincides with the global maximum in conformational heterogeneity of NTL9 (fig. 7.14 and figure 4B of Lyle *et al.* [348]), suggesting that for ensembles of finite polymers, heterogeneity may be a useful measure for considering the formation of local and long-range order in the unfolded states.

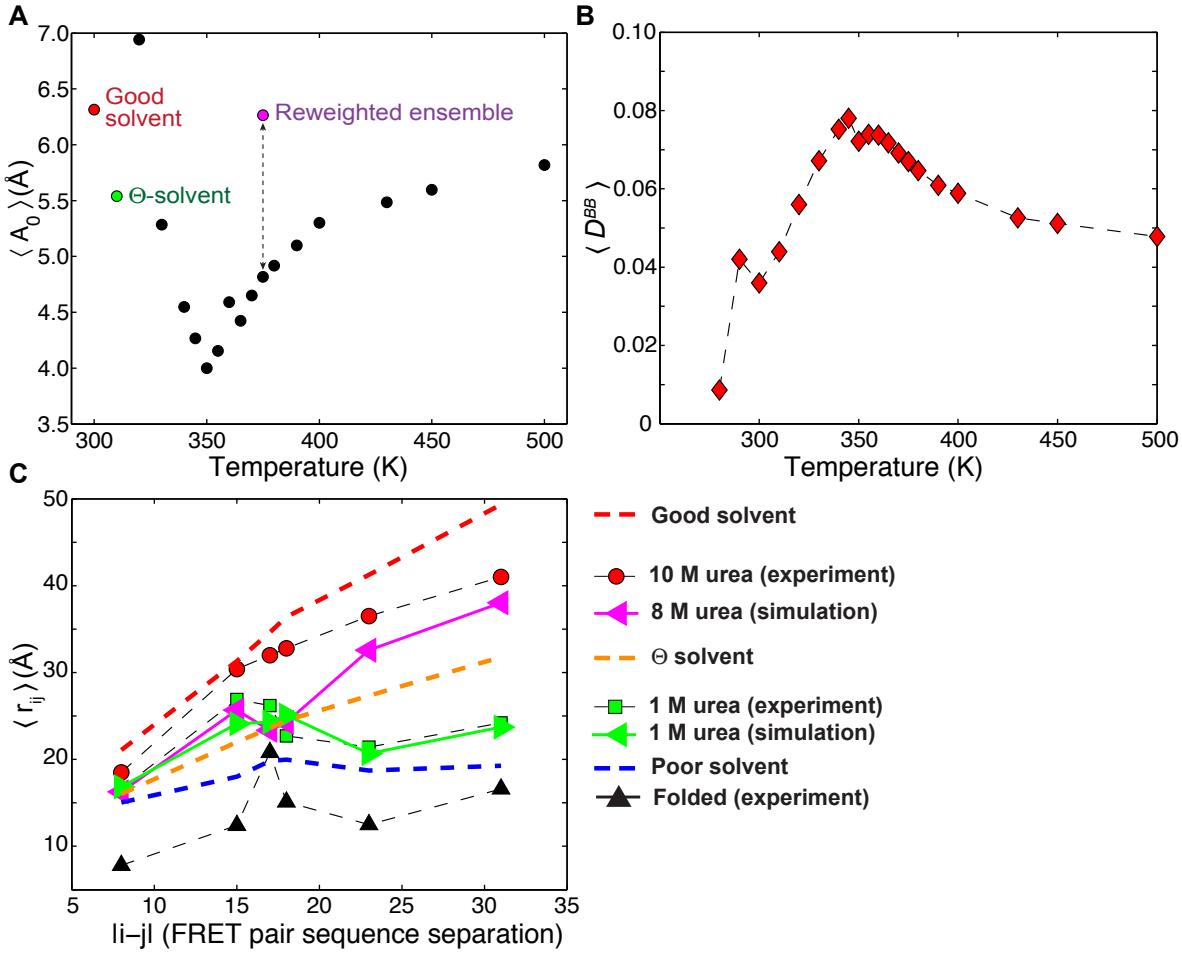

Figure 7.14: We quantified A_0 based on our internal scaling fits, and found anon-monotonic relationship in which A_0 decreases from the athermal limit towards ~350 K, at which point A_0 rapid increases again (panel A). A_0 can be considered a combination of the average volume occupied by a monomer and the chain persistence length. We suggest that the inflection point may reflect the onset of stable structure formation. A similar peak is observed when examining the backbone-derived heterogeneity; the most heterogeneous ensembles also have the lowest A_0 value (panel B). Panel C provides a summary of the different scaling behaviours for real data, simulations, and theoretical limit behaviour.

7.4 Discussion

In this work, we offer a high-resolution description of the unfolded state under native conditions obtained through a combination of multiple novel and non-invasive FRET probes coupled with time resolved FRET measurements, time resolved SAXS, extensive all atom simulations, and polymer theory. Taken together, our results suggest the unfolded state more compact than the fully denatured state yet still relatively expanded, conformationally heterogeneous, and engages in the formation of extensive secondary and tertiary structure elements of both the native and non-native variety.

7.4.1 NTL9 Obtains Consistent Results Between SAXS and FRET

These results are in good agreement with recent computational studies of protein L, where a modest compaction in the unfolded state is observed before folding, and of a 17% contraction in the unfolded state of ubiquitin upon dilution from 6 M to 0 M GdmCl [354,477]. Native and non-native interactions that preceding folding and are driven by hydrophobic residues have been identified in the early stages of monellin folding, consistent with a model of structure formation in which hydrophobic residues pre-organize the unfolded state to facilitate folding [11, 12, 48, 204, 655]. The results are also fully in line with a series of recent papers examining the unfolded behaviour of proteins, and in agreement with earlier work on protein folding [20, 59, 468, 673].

These studies highlight the utility of minimally perturbative unnatural amino acids as probes of protein structure. We expect that the FCN-Trp pair will facilitate studies of single domain protein dynamics or subdomains within larger proteins. Stability measurements demonstrate

that FCN is a relatively non-perturbative substitution in NTL9, not only for Tyr and Phe, but also for Lys and Gln. The attachment of larger fluorescent dyes to single domain proteins has led to arguments that the fluorophores may be inducing the chain to collapse. These studies demonstrate that contraction is observed even using the least perturbative probes.

7.4.2 Extensive Conformational Heterogeneity Emerges Under Native Conditions

The conformational heterogeneity associated with unfolded proteins, especially in the unfolded state, is shown by the non-monotonic relationship between sequence and spatial separation in fig. 7.8. This behaviour highlights the importance of probing multiple regions within the chain to obtain a full description of the ensemble. Moreover, the relatively short spatial distance between probes reduces the 'extrapolation' provided by polymer models, as demonstrated by the fact that both the WLC with diffusion and the Gaussian distribution yield similar distances. When probe distances are large the gamut of conformational behaviour that could occur in the intervening space is extremely likely to be underestimated by even sophisticated polymer models. By combining multiple FRET pairs over a relatively short distance the uncertainty associated with the FRET-to-distance conversion is minimized.

The unfolded state of NTL9 contains both native and nonnative elements of structure. The 1 M urea ensemble contains residual native helical structure in regions corresponding to the first and second α-helix. Previous NMR experiments performed on the destabilized double mutant, V3A-I4A, have also detected residual helical structure in the absence of denaturant for both helices, however, the extent was greater for the second rather than the first,

C-terminal α-helix [376]. Experiments performed on peptide fragments, two of which corresponded to the first and second α-helix, have shown that the second helix partially forms in isolation while the first does not [308]. Helical structure in the C-terminal helix thus requires only local contacts to form, while the first helix may be stabilized by tertiary interactions. In the NMR studies using V3A/I4A, the truncation of two hydrophobic residues, V3 and I4, which form hydrophobic clusters with residues in regions of the first helix, may act to destabilize the first α-helix in V3A-I4A. These residues may participate in long-range interactions in the wildtype protein, and residual helical structure in the region of the first helix may be more strongly stabilized.

Previous results indicated the presence of both native and non-native contacts in the unfolded ensemble in 8 M urea, and both native and non-native contacts are observed in the 1 M ensemble in the present study. Indirect experiments probing the unfolded state of NTL9 have shown that lysine-12 forms non-native electrostatic interactions in the unfolded state, and that this residue is coupled to the formation of hydrophobic clusters that are distant from K12 both in primary sequence and spatially in the folded structure [99]. NMR PRE experiments combined with Monte Carlo simulations have provided evidence for hydrophobic cluster formation in the 8 M urea unfolded ensemble [377]. Many of the residues involved in these clusters are conserved among different NTL9 sequences. It has been proposed that hydrophobic clusters of Leu, Ile, Val and Phe occur in the unfolded state because they are formed early in folding [655].

Several roles for non-native contacts in the unfolded state have been proposed. Hydrophobic clustering may reduce the likelihood of aggregation that may be initiated from exposed hydrophobic residues. Studies have reported that certain non-native interactions can lower the free energy barrier for folding and increase the folding rate, and that non-native contacts

can act to constrain the formation of native contacts [42, 84, 108, 168, 458]. Other studies have argued that non-native contacts do not usually play a role in the folding mechanism, but may nonetheless influence folding rates and modulate free energy landscapes [45, 201].

At first glance the FRET and SAXS results may seem at odds. The FRET data indicates significant collapse of the unfolded state under folding conditions, with some segments experiencing up to 40% compaction. The SAXS data indicates a more modest ∼20% contraction, and examination of the Kratky profile indicates that the unfolded state does not assume an overall globular conformation. Monte Carlo simulations, however, show that the FRET and SAXS results are consistent. The unfolded state is heterogeneous; certain regions of the protein experience significant collapse while others remain relatively expanded. The overall ensemble-average dimensions resemble that of a polymer in a Θ-solvent. Such global properties have been reported for several other proteins that were studied using single-molecule FRET and/or SAXS [20,59,628,673]. The previous FRET studies only examined a construct with labels at the N and C-termini, and was thus unable to probe local conformational preferences and distinguish heterogeneous collapse from uniform collapse. SAXS measurements can only report on overall chain properties. The fact that the global dimensions of the 1 M urea unfolded ensemble of NTL9 resemble a chain in a Θ-solvent does not mean that the unfolded state behaves as an ideal chain, devoid of long range conformational preferences. On the contrary, we provide direct evidence for heterogeneity within the ensemble and the presence of significant elements of structure. The ensemble consists of molecules that rapidly fluctuate between compact and coil-like conformations.

The interconversion between these two states may facilitate the search for the folded state more effectively than an expanded coil-like or compact globular ensemble. In an expanded random coil-like state, there may be little or no bias to initiate the folding process. In a

compact globular state the rate of reconfiguration of the chain may be significantly reduced because of internal friction and this could slow down the search for the folded state [108, 458]. Taken together, our results are consistent with an emerging picture of the unfolded state under folding conditions behaving in an relatively expanded yet heterogeneous manner, with nascent structure flickering into and out of existence, simultaneously facilitating the nucleation of folding while efficiently exploring conformational space.

Chapter 8

'Resolving' the Controversy Between SAXS and FRET

The following chapter includes ideas taken from the paper **SAXS vs. FRET: A Matter of Heterogeneity** by K.M. Ruff and A.S. Holehouse, to be published in the *Biophysical Journal* in September 2017. It also includes ideas from the manuscript **Collapse Transitions of Proteins and the Interplay Amongst Backbone, Sidechain, and Solvent Interactions** by A.S. Holehouse and R.V. Pappu, which is currently under review for *Annual Reviews in Biophysics*, with an expected publication date of May 2018. All analysis and text presented here was generated by A.S.H. The initial version of the dye-addition approach (that has become COCOFRET) was developed by K.M.R. .

8.1 Background

The average properties of the denatured state under strongly denaturing conditions are accurately described as a polymer in a good solvent, as demonstrated in seminal work by

Wilkins *et al.* and by Kohn *et al.*, and discussed previously in chapter 5 [297,641]. In contrast, a molecular description of the unfolded state under low denaturant and/or folding conditions has remained less well understood. Historically, it has been argued that the unfolded state under low denaturant conditions exists as a fully expanded self-avoiding random walk, as a compact but disordered globule, or as polymer in a Θ-solvent [20,59,234,255,538,673]. These divergent descriptions originate in part from apparently contradictory results from Small Angle X-ray Scattering (SAXS) and single molecule Förster Resonance Energy Transfer (smFRET) experiments.

The discrepancy is summarized schematically in fig. 8.1. SAXS results have repeatedly shown that as denaturant concentration is diluted the global dimensions of unfolded proteins remain largely invariant (within some error of $\sim$ 1-3 Å) before a precipitous drop in global dimensions during folding at low denaturant concentration [128,255,457,550,628,661]. In contrast, smFRET studies have shown a continuous decrease in global dimensions as the denaturant concentration is lowered [20, 59, 379, 673]. The smFRET result suggests that intermediate states between the fully denatured and folded endpoints exist across a continuum of global dimensions.

Recently, there have been several reports providing insight into the molecular origins of this discrepancy. Elegant work from the Reddy group suggests FRET over-estimates of the radius of gyration (R_G) at high denaturant concentrations, artificially enhancing the apparent extent of collapse [354]. Work from the Grishaev, Best, and Schuler groups provides a rigorous molecular dissection of several proteins using four different experimental approaches, and converge on the discrepancy originating from a combination of factors, including the inherent sensitivity of FRET, the sensitivity of SAXS to the fitting range used for the Guinier analysis, and challenges associated with determining the radius of gyration at low

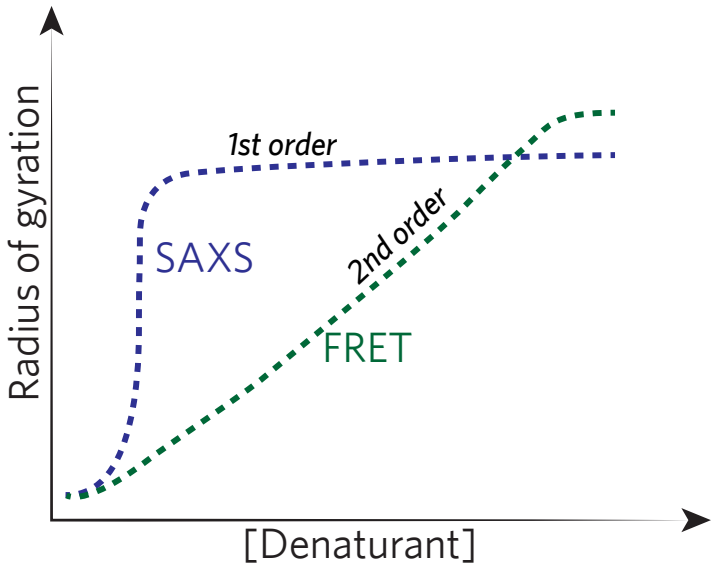

Figure 8.1: Stylized schematic showing the difference between SAXS and FRET. SAXS results have typically reported that chain dimensions are broadly insensitive to denaturant concentration up to some threshold value, at which point the protein collapses and folds. FRET results have reported that chain dimensions follow broadly follow a continuous response to denaturant concentration.

denaturant concentrations [59,673]. A computational study by Reddy and Thirumalai offers similar conclusions, suggesting the formation of secondary structure elements occur in the Θ-state and precedes full folding [477].

In chapter 7 we described the conformational behaviour in the unfolded state under folding conditions of the N-terminal domain of the ribosomal L9 (NTL9) protein. To briefly summarize those results, we identified well defined sequence-encoded structural preferences in the unfolded state under folding conditions. These structural preferences give rise to local deviations in conformational behaviour from that expected of a polymer in a Θ-solvent, despite the fact that the globally averaged values are indistinguishable from this theoretical

limit. This result reconciles the apparently contradictory results that the unfolded state under folding conditions can act as a crucible for folded structure (nucleation) while remaining highly expanded; for a complex hetero-polymer such as a polypeptide, the simultaneous acquisition of transient structure coupled to an expanded state appears to be an emergent property of the amino acid sequence. In contrast, such behaviour is simply not realizable for a homopolymer.

These results led us to wonder if the deviation between SAXS and smFRET at low concentrations of denaturants could, rather than cause for concern, be a hallmark of the formation of local and/or long range structure in the unfolded state, as we had observed in NTL9. The conversion of FRET efficiencies to distances relies on the use of polymer models. This approach has been enormously powerful, but makes strong limiting assumptions which in the most extreme cases require the use of entirely unrealistic parameters to fit experimental data, indicative of failures in limiting assumptions made in the application of those models. At high concentrations of denaturant unfolded proteins show behaviour is in good agreement with standard polymer models. However, as denaturant concentration decreases and transient long-range and local structure begins to form, the conformational behaviour of the chain will increasingly deviate from conventional polymer models and sequence-specific local and long range interactions begin to play an increasingly important role in determining the conformational ensemble. We propose that as denaturant concentration is lowered, the reduction in solvent quality leads to a precipitous drop in the ability of homopolymer models to accurately describe heteropolymeric ensembles. In effect, at high denaturant concentrations the linear chemical heterogeneity encoded by a polypeptide becomes irrelevant and chain conformation properties are dominated by highly favourable chain-solvent interactions, yielding a psuedo-homopolymer. As denaturant concentrations decrease, this chemical heterogeneity has an increasing impact, and the polymer's conformational behaviour transitions from a

psuedo-homopolymer into a heteropolymer. Consequently, this leads to a precipitous drop in the ability of homopolymer models to accurately capture the conformational behaviour in a meaningful way, as sequence specific effects introduce anisotropic deviations from ideal chain behaviour.

8.2 Methods

To test this hypothesis, we examined the unfolded state ensemble of protein L. Protein L has been extensively studied by SAXS and smFRET under a variety of denaturing conditions. We generated an unfolded-state ensemble at all-atom resolution using the CAMPARI Monte Carlo simulation engine and ABSINTH implicit solvent model as described previously. Specifically, starting from a random coil conformation we generated an extensive unfolded ensemble at 375 K. We then used a post-processing approach to add ensembles of Alexa488 and Alex594 dyes to the terminal residue of each protein conformation and calculated the converged mean inter dye distance (see chapter 9 for a full description of this process). Fig. 8.2 provides a structural rendering of what a single conformation with 'clouds' of dyes at each termini look like. The ensemble average FRET-derived distance is determined by computing the FRET derived distance between each unique pair of dyes and computing the average. This procedure is repeated for every protein conformation to construct an average FRET-derived distance that is an ensemble average of ensemble averages.

We next defined a set of target distances for the dye-dye distances, and used a Maximum Entropy and χ^2 based re-weighting procedure to re-weight the ensemble to match specific target dye-dye distances (see methods in 7 for further details). For each re-weighted ensemble, we then calculated the radius of gyration (R_G) using the relation $R_G^{EE} = \sqrt{\frac{R_{EE}^2}{6}}$ and

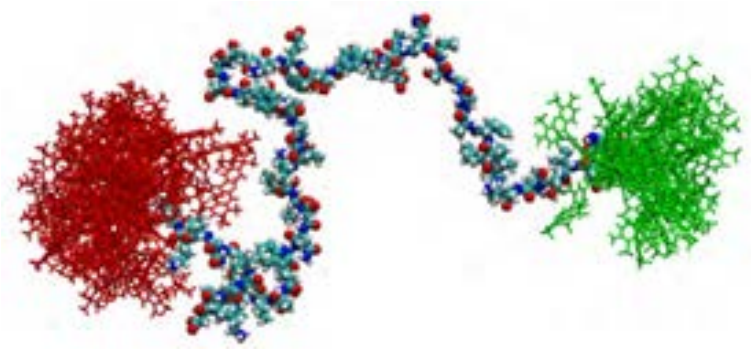

Figure 8.2: Structural representation of the Alexa488 and Alexa594 dyes on a single protein L conformation. Each dye ensemble contains several hundred unique pairs of dye conformations.

the definitive R_G^{GEO} using the full set of atomic coordinates. R_G^{EE} is equivalent to the radius of gyration calculated from smFRET experiments, while R_G^{GEO} is equivalent to the radius of gyration from SAXS.

To recast this more simply, the approach can be thought of as using the smFRET-derived result of a continuous transition to bias the derived ensemble, and then ask how the radius of gyration derived from these biased ensembles calculated in two different ways behave.

8.3 Results and Discussion

The results from this analysis are shown in fig. 8.3

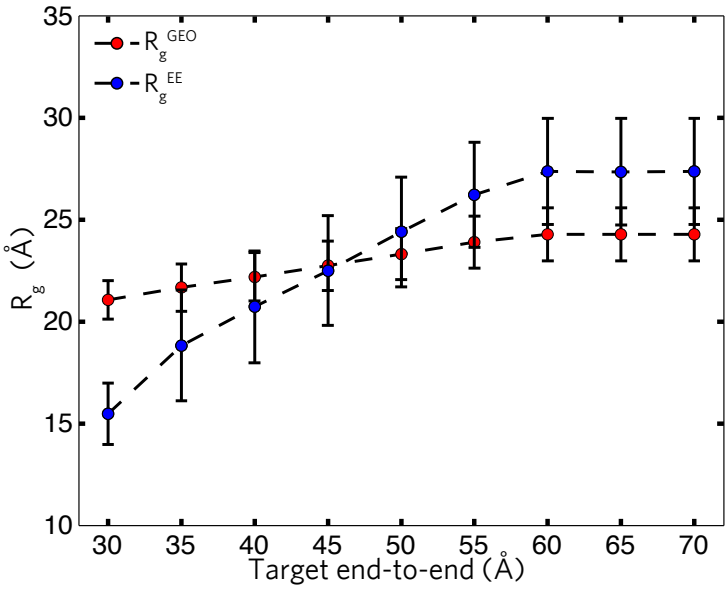

Figure 8.3: Results from analysis of re-weighted ensembles. R_G^{GEO} as calculated using all atomic positions, changes by only $\sim$ 2 Å , while R_G^{EE} as back-calculated based on dye-dye distances, changes by $\sim$ 12 Å

The results obtained for this approach are quantitatively consistent with the results observed in experiment. The R_G^{GEO} is largely insensitive to the conformational perturbation induced by the re-weighting procedure, while the R_G^{EE} changes by around 50%. This ignores any additional influence of different analysis approaches, perturbation to the ensemble due to dyes, inherent technique bias, or any other previously offered explanation, which we readily agree may additional contribute to the observed differences between the methods. Instead, our results invoke only simple polymer behaviour and conformational biases to reproduce the observed discrepancy. Importantly, the discrepancy directly observed here is larger than

the differences observed in real systems (which typically vary between 20-40%), hence while this may be an extreme example, we suggest that this can account for a significant fraction of the apparent discrepancy between SAXS and FRET.

To determine if this result is specific to protein L or a more general phenomenon, we repeated this approach using smFRET and SAXS results examining the unfolded state of ubiquitin at pH 2.5. Dye-dye RMSD (including linkers) for N and C terminal dyes were determined from transfer efficiency histograms at 2 M, 4 M, 6 M and 8 M urea to be 7.7 nm, 7.0 nm, 6.3 nm, and 5.4 nm, respectively [20]. Using these values as dye-dye distance targets, we re-weighted a denatured state ensemble of ubiquitin and calculated the derived R_G values using both methods. The results of this analysis are shown in fig. 8.4.

As with protein L, R_G^{GEO} changed by $\sim$ 2.5Å, while the R_G^{EE} changes by $\sim$ 7.5Å. These values are fully consistent with experimentally derived SAXS values, which at 8 M urea were found to be $28\pm, 3.5$Å. In summary, the $R_G^{GEO}$is statistically unchanged in the re-weighted ensembles, while the end-to-end distance-derived R_G^{EE} shows a continuous transition.

Fundamentally, SAXS and smFRET report on different order parameters. The assumption that they should match one another relies on the use of limiting case statistical polymer models which may or may not be appropriate. These models work well in the fully denatured state, where local and long-range conformational biases are minimal (though still present). As denaturant concentration decreases and sequence encoded conformational preferences become stronger one would expect these models to become less relevant. In short, given the sequence dependence associated with intrinsically disordered proteins (IDPs), and a growing body of evidence describing the formation of native and non-native structure in the unfolded state, there should be no a priori expectation that SAXS and smFRET should agree at low concentrations of denaturant.

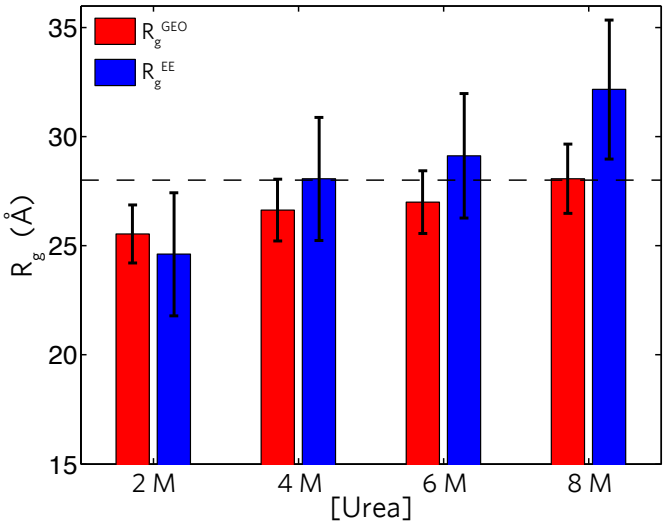

Figure 8.4: Results from analysis of re-weighted ensembles for ubiquitin. Dashed line shows the SAXS derived radius of gyration obtained at pH 2.5 and 8 M urea. Note the agreement between the SAXS derived result at 8 M urea and the R_G^{GEO} at 8 M urea is an emergent property of the ensemble and not something prescribed in the reweighting, giving us confidence this approach represents a reasonable method for the reconciliation of SAXS and FRET results.

There are two important implications from this result. The first is that the generalization of this idea will strongly depend on the sequence of interest and its conformational propensities in the unfolded state. There may be proteins where SAXS and FRET show good agreement, suggesting the absence of strongly anisotropic conformational preferences. At lower denaturant concentrations there may also be sharp transitions in smFRET derived distances as locally cooperative units fold. A critical takeaway is that no two sequences will necessarily behave the same way, although we tentatively suggest that the prevalence of the SAXS/smFRET discrepancy could be considered evidence for well-defined conformational preferences in the unfolded state as a general feature of foldable proteins.

The second implication is that the results from SAXS and smFRET experiments provide critical and complementary information. Recent work appears to largely lay to rest the notion that proteins in general undergo either rapid collapse to a globule before folding, or that folding occurs as a two state phenomena from expanded self-avoiding random walk to folded state. However, there does appears to be a modest global contraction of chain dimensions upon dilution of denaturant, which occurs concomitantly with specific conformational biases. In the absence of other techniques, SAXS would be blind to the extent of these conformational biases, but an extrapolated R_G based solely on end-labelled dye pairs would suggest global dimensions drop systematically and precipitously as denaturant concentrations are diluted. Instead, taken together, they paint a complex picture of the unfolded state, which aligns well with a growing consensus that the unfolded state under folding conditions behaves like a polymer in an effective Θ-solvent, albeit with anisotropic conformational preferences.

The most comprehensive approach for obtaining an accurate description of the polymeric properties of an unfolded protein is the use of multiple FRET pairs within the same sequence, as described in chapter 9. This can be expensive and labour intensive, but especially when combined with other techniques such as SAXS, NMR, and simulations offers a near unequivocal description of the solution behaviour of unfolded proteins. Finally, the ideas presented here translate directly to IDPs, where despite their 'unstructured' nature, well defined conformational preferences can be manifest through local and long range interactions leading to deviation between limiting models and real chain behaviour as demonstrated explicitly in recent work.

Chapter 9

CTraj: An Analysis Framework for All-Atom Simulations of Disordered Proteins

The following section is taken from a manuscript in preparation with the working title **CTraj: an analysis framework for all-atom simulations of disordered proteins** by A.S. Holehouse, K.M. Ruff, J. Lalmansingh, N. Lyle, A. Vitalis, and R.V. Pappu. All CTraj code, was written by A.S.H. The entire FRET fitting procedure was conceived of, developed and originally deployed by K.M.R, with COCOFRET providing the convergence filter, a higher throughput, and more general implementation. J. Lalmansingh was involved in software testing. N.L. developed and implemented the original PRE analysis code. A.V. developed many of the original analysis tools which have been ported to CTraj.

9.1 Background

For many of the projects in this work, we may wish to analyse a variety of properties associated with all-atom simulations of unfolded proteins. While various simulation analysis packages exist, these are typically based around folded proteins and as such contain a set of analytical tools appropriate for structural characterization but less useful for describing ensembles of unfolded proteins [374, 384]. Conversely, the types of order parameters that are critical for characterizing a disordered protein's ensemble (internal scaling, end-to-end distance, asphericity, radius of gyration *etc.*) are likely to be relatively uninformative when applied to a simulation of a folded proteins. While various analysis routines are built directly into CAMPARI, developing new bespoke analysis in FORTRAN is a challenging process due to a general lack of modern libraries, an inability for interactive (e.g. command line) testing, and the associated compile time.

To alleviate this issue, we have developed a simulation analysis framework specifically for analysing CAMPARI simulations of disordered proteins. CTraj is written in the Python programming language, and was developed with flexibility in mind. It takes advantage of the underlying robustness of the MDTraj analysis toolkit to parse trajectory files of a wide variety of types (`pdb`, `xtc`, `dcd`, `netcd`, `HDF5`) but provides a suite of novel analysis routines to describe conformational ensembles of disordered proteins. CTraj has been used extensively within the lab, is undergoing private beta testing by several colleagues, and is currently on version 0.2.14 (with version 0.1.0 released in May 2015). The codebase is over 10,000 lines of code, and has native implementations of a wide range of simulation analysis protocols. The majority of analysis routines can also accept a vector of trajectory frame weights, allowing re-weighted ensembles (based on COPER or T-WHAM) to be analysed directly [102, 324]. While developed specifically for CAMPARI trajectories, CTraj has been used without issue

for simulations performed with both GROMACS and NAMD, and is being tested by several users outside of the lab.

The remainder of this short chapter is set up as follows. We provide an overview of the functions provided by CTraj in the form of an infographic (fig. 9.1). We then outline the analysis approaches that are either unique to CTraj (in as much as they are not implemented in other analysis packages, as far as we know) or an entirely novel algorithm developed and implemented in CTraj.

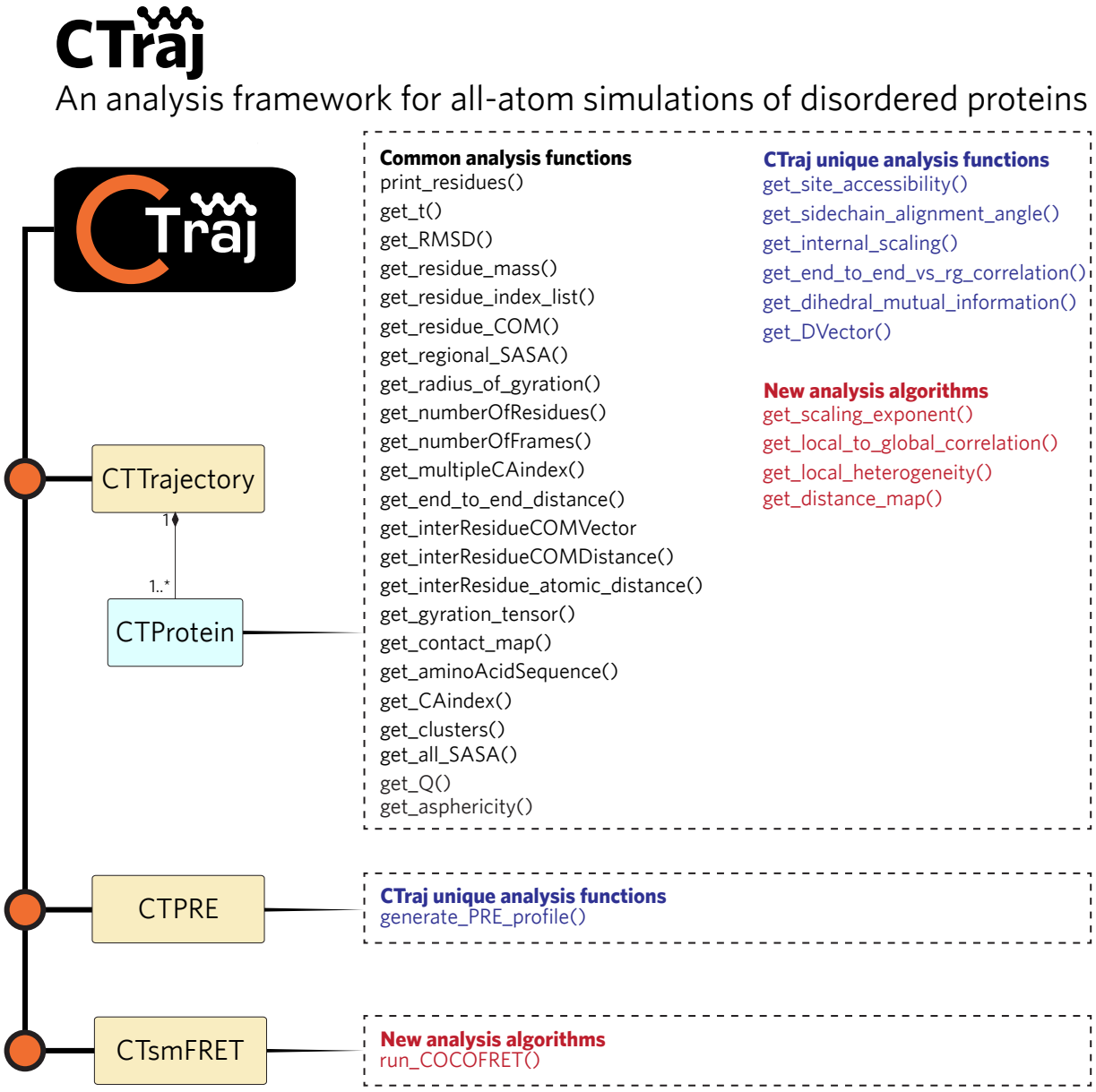

Figure 9.1: Schematic overview of the software architecture associated with CTraj. Upon reading a trajectory in one or more CTProtein objects are generated (one per polypeptide chain) and various protein-specific analysis can be performed though these CTProtein objects. CTProtein objects can also be passed to CTPRE and CTsmFRET objects for PRE and FRET based analysis.

9.2 Methods

CTraj is written in Python programming language. The underlying input and model representation used in CTraj is built on on MDTraj (version 1.8.0 or higher) [374]. MDTraj is a Python-based molecular dynamics analysis framework that combines the convenience of Python with the robustness of a well developed trajectory I/O library. The trajectory I/O components of MDTraj are built on the OpenMM trajectory readers, affording extremely robust and efficient and reliable input and output. Analysis tools are built using the Numpy (version 1.12.0) or higher and SciPy libraries. Version control is provided by GitHub.

9.2.1 CTraj Unique Analysis Functions

The following represents analysis approaches used for disordered proteins but are not available out-of-the-box in other analysis packages (to the best of our knowledge).

`get_site_accessibility()`: This method allows the user to determine the solvent accessibility of a given residue (or all residues of a particular type) for a probe of a given size. This is convenient for quantifying the accessibility of (for example) hydrophobic residues across a sequence without needing specifying which hydrophobic residues are of particular interest. `get_sidechain_alignment_angle()`: This method allows the user to quantify the degree of alignment between a pair of sidechains. The sidechain vector is by default defined as the C^{α} atom to a sidechain specific 'master' atom, where the master atom is sidechain specific. The identity of this master atom can also be changed via an input argument, as a string corresponding to the atom name. This method is useful for identifying correlations in sidechain directions in an ensemble. `get_internal_scaling()`: creates an internal

scaling vector for the sequence. See chapter 5 for a detailed discussion on internal scaling profiles. Either center-of-mass internal scaling or C^{α} internal scaling can be performed. `get_end_to_end_vs_rg_correlation()`: function that provides a formal assessment of how well the end-to-end distances correlates with the true radius of gyration. In a true Gaussian chain these two properties are directly related to one another by a functional form, such that while the absolute values may vary depending on the the apparent solvent quality the correlation in terms of this form can be computed. Useful for assessing how well a Gaussian chain approximation captures the conformation behaviour observed by the chain. `get_dihedral_mutual_information()`: automatically compute the mutual information between a pair of dihedral angles (backbone or sidechain). `get_DVector()`: computes the full $\mathcal{D}$ vector, as defined by Lyle *et al.* [348]. The $\mathcal{D}$ vector provides a description of how similar the combined set of conformations that make up the ensemble are to one another, and the mean value is a reasonable (albeit un-normalized) proxy for the global heterogeneity associated with the ensemble.

9.2.2 New Analysis Algorithms

The following represents entirely new analysis algorithms developed within CTraj.

get_scaling_exponent()

This is a novel algorithm that attempts to fit a subsection of the internal scaling data to extract the apparent scaling exponent ν^{app}. For a discussion on the meaning of ν^{app} see chapters 2 or 5. Various iterations of this function have been used, with the default behaviour now fitting the expression:

$$\sqrt{\langle\langle r_{i,j}^2\rangle\rangle} = A_0|i-j|^{\nu_{app}} \tag{9.1}$$

The following two corrections for finite-chain behaviour are also made by default, although both corrections can be modified to explore how finite-chain corrections influence the derived ν^{app}.

1. We use a threshold of $|i-j| > 15$, such that only pairwise distances used are those where i and j are 16 residues or more apart. This threshold of 16 was predicted based on the fact that sequence-specific fractal scaling behaviour is not expected to occur on length-scales below the blob-size (5-7 residues), and as such a minimum of two blob-lengths should be necessary to determine scaling behaviour. Empirically, we have found that this threshold allows the robust reproduction of various theoretical limits.

2. We discard the five largest sequence separations to avoid a biasing influence of the largest separations, which experience the least double averaging and are the most sensitive to heteropolymeric deviations (i.e. $|i-j| < (N_{res}-5)$). While these deviations are real, our goal in analyzing the scaling exponent is to identify a mean-field parameter that captures the apparent scaling exponent, and heteropolymeric deviations should be examined and explored using other approaches

Fitting is done in the true domain (as opposed to the log domain) to ensure all sequence separations are equally weighted. We have previously analysed the 2D surface of ν^{app} vs. A_0, as shown in fig. 9.2; in all cases tested this surface has been convex, with a well defined minimum, suggesting there is always a best-fitting ν^{app}. As ν approaches 1/3 the fractal assumption becomes less valid. A best possible scaling exponent can still be fit, but whether

or not the internal scaling data shows a good linear fit is a separate question. As a result, the apparent scaling exponent is a convenient classification tool, but should be treated as a qualitative factor who's meaning will vary depending on the specific amino acid sequence in question. In chapter 7 we used this method to evaluate how the scaling exponent and A_0 prefactor change for unfolded ensembles as a function of temperature.

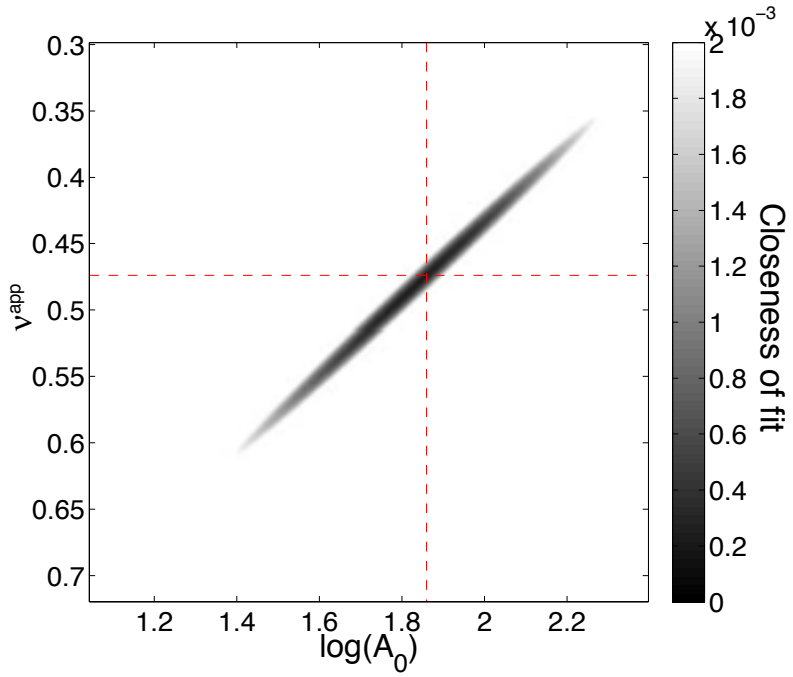

Figure 9.2: The darker the colour the better the fit. Red dished lines show the intersection the global minimum. Note that a range of values fit the data approximately equally well.

get_local_to_global_correlation()

To what extent do global and local distances correlate with one another? This algorithm asks "If n pairs of residues are selected at random and used to determine the expected global dimensions for the ensemble, how good is that prediction?". This question is asked for

an asymptotically large set of randomly selected pairs (where n pairs without replacement are randomly selected in each iteration) until the average predictive power that n pairwise distances will provide has been determined. This ensemble-average pairwise predictive power is referred to as $\langle \xi_n \rangle$, where n is the number of pairs. This approach can be performed for several different n. As an example, this analysis was performed as a function of temperature on different ensembles of the unfolded state of NTL9 (see chapter 7 for further discussion). For lower temperature/folded ensembles even five pairs of residues is entirely insufficient to accurately predict global dimensions, but for increasingly heterogeneous ensembles (i.e. at higher temperature) three or more pairs is (on average) sufficient to - with reasonable accuracy - predict the global dimensions. For reference even for a 'perfect' ensemble in the Θ-state (e.g. the Gaussian chain) the $\langle \xi_n \rangle$ value obtained with five pairs is $\sim$ 0.8. The inability to reach 1.0 reflects the fact that this approach is determining the *average* correlation for n pairs picked randomly, such that in the process of selecting n pairs there is a chance of randomly picking 2 residues next to one another, and indeed of doing so multiple times. An alternative approach would be to describe the distribution of correlations, rather than the ensemble average correlation. This approach could ultimately be modified and used to aid in selecting the most informative positions to placed FRET labels based on results from all-atom simulations.

get_local_heterogeneity()

As discussed in chapter 2 assessing the quality of sampling (conformational heterogeneity) is critical to assess if a simulation of an IDP is truly describing an conformationally heterogeneous ensemble, or if we are simply describing a rigid body with sidechain re-arrangements. To compute local heterogeneity, the sequence is subdivided into overlapping ten-residue

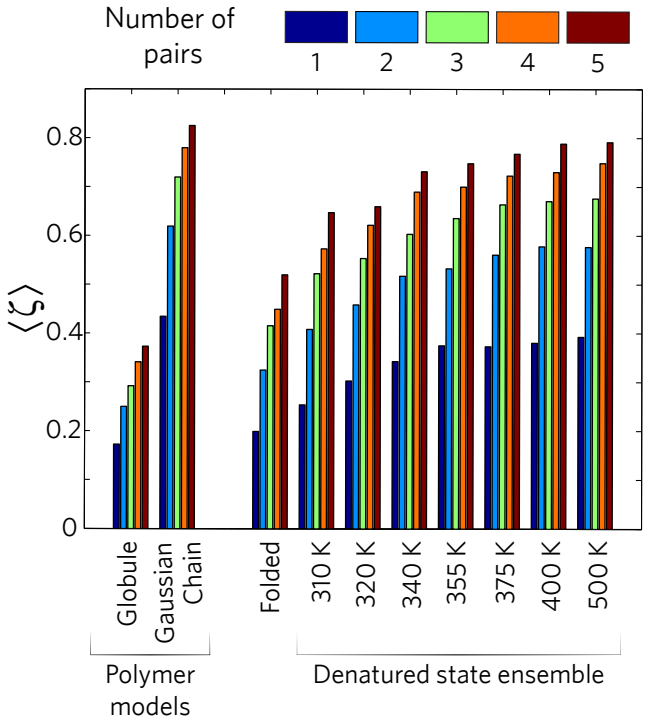

Figure 9.3: Comparison of how $\langle \xi_n \rangle$ changes as a function of number of pairs and temperature.

fragments and the all-vs.-all RMSD for each of those fragments is computed, generating a distribution of RMSD values for *each* overlapping fragment. This set of distributions can be plotted as a function of fragment center to generate 2D heatmap that quantifies local conformational heterogeneity in an intuitive and non-parametric way. As an example, see the fig. 9.4. In this analysis we quantified the local heterogeneity associated with the protein α-synuclein. The N-terminal and C-terminal regions are highly heterogeneous, while the central hydrophobic region (and repeats 2 and 3) are much less well sampled, suggesting they typically fall into local meta-stable minima. This approach provides a useful tool for assessing how well sampled a simulation of disordered protein is, and can identify local regions that may not experience conformational sampling without relying on visual inspection.

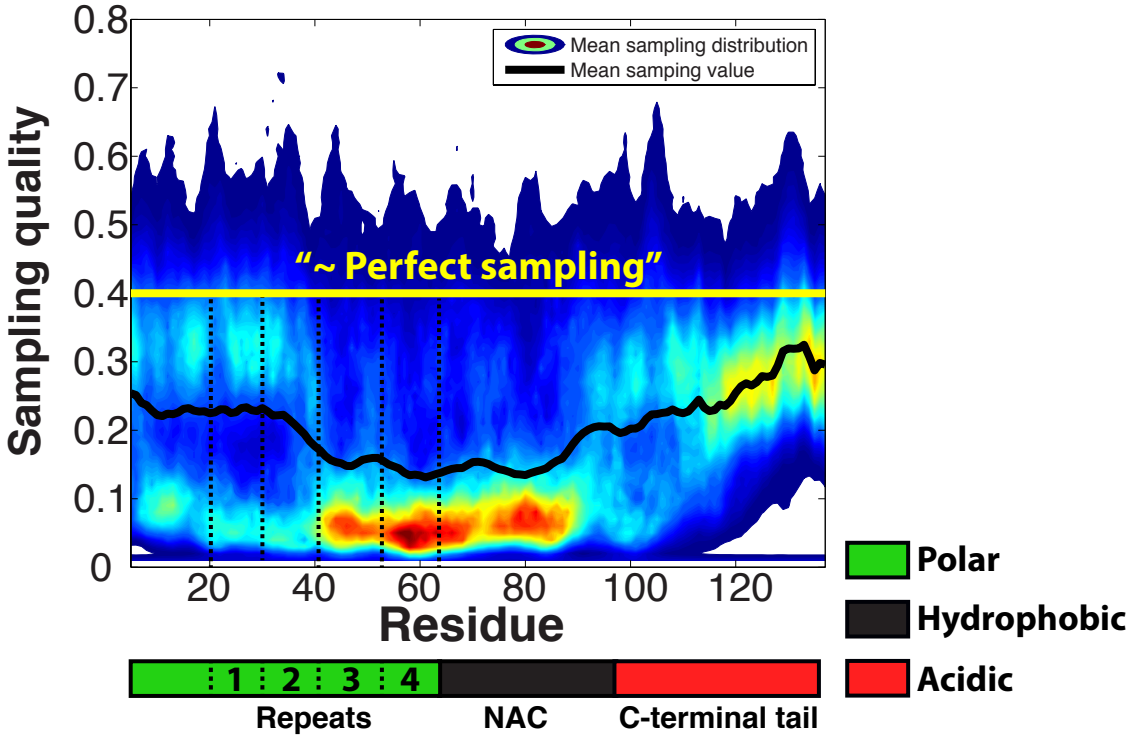

Figure 9.4: Visual representation of local conformational heterogeneity described in terms of sampling quality. A low value of sampling quality suggests a small number of structurally distinct conformations are explored during the simulation, indicative of locally trapped states. For a folded protein this sampling quality would apparently be different, but the local structure would be consistent across many independent replicas. For a rugged energy landscape with many deep local minima, independent simulations would yield entirely different locally trapped states. The perfect sampling line describes the local sampling experienced by a Flory Random Coil, which represents the ensemble of maximum heterogeneity [348]. In our example the N-terminal third of the ensemble shows a bimodal distribution, suggesting this region experiences both disordered and heterogeneous conformational behaviour and transient local interactions leading to meta-stable states.

get_distance_map()

We have used scaling maps extensively in throughout the chapters in work (see chapters 7, 6, 11). Scaling maps are generating by normalizing a distance map taken from a full simulation by a distance map generated from a sequence-matched EV simulation (see 2 for a discussion on the EV ensemble). Scaling maps provide an intuitive and two-dimensional representation of the average conformational preferences associated with the polypeptide, allowing us to discern local and global conformational properties. For more information on scaling maps see section 6.3.8.

9.2.3 CTPRE

CTPRE is a separate class which can be passed a CTProtein object. From this an ensemble specific paramagnetic resonance enhancement (PRE) profile can be generated. Briefly, PRE is an experimental technique whereby a paramagnetic spin label (typically nitroxide) is chemically attached to a specific residue (typically a cystein). The proximity of each residue to this spin label influences the rate of spin relaxation which in turn can be monitored directly by NMR, such that residue-by-residue profiles can be generated whereby the rate of relaxation is related to the distance from the spin label. This provides an experimental approach to determine the relative distance between certain residues. The relationship between distance and the spin label is complex and depends on many factors. In work by Meng & Lyle *et al.*, the authors use this approach to generate PRE profiles from NMR, and equivalent profiles from simulations [377]. The method implemented here reproduces that approach. However, it does not take local shielding, spin-label behaviour, or solution effects into consideration, all of which represent future enhancements that we would like to

implement going forward. Recent analysis suggests that for more complex IDPs these factors must be taken into consideration, and improvements to this function to do this are under way.

9.2.4 CTsmFRET

CTsmFRET provides an out-of-the-box implementation of the COCOFRET algorithm (COnformational COnvergence). In brief, COCOFRET is a method of building a cloud of dye positions onto each protein conformation of an ensemble (with dyes placed at specific positions in the amino acid sequence), and computing the mean converged FRET efficiency associated with that cloud, taking into account the dye-dye distance and the relative orientation. The approach involves building an ensemble of dye conformations at each position independently via a Monte Carlo trial and reject approach. Dye conformations are selected from the Handy-FRET library (http://karri.anu.edu.au/handy/), ideal bond angles between the peptide and dye are fit, and steric overlap is provided in conjunction with a mean-field hydration shell. Having placed as many dye rotomers at each position after making n attempts, the complete set of cross-dye FRET efficiencies are computed by determining the dye-dye distance, calculating an orientational dependent κ[25] value, and using that to define the R_0 (see subsection 2.3.3 for additional discussion on FRET. We define κ in the usual way:

$$\kappa = \hat{\mu_D} \cdot \hat{\mu_A} - 3(\hat{r} \cdot \hat{\mu_A})(\hat{r} \cdot \hat{\mu_D}) \tag{9.2}$$

[25]NOTE that κ here is not related to charge patterning!

Where $\hat{\mu_D}$ and $\hat{\mu_A}$ are the unit vectors along the transition dipoles of the donor and acceptor fluorophore and $\hat{r}$ represents the unit vector associated with the dye-to-dye distance. Note that κ is a *purely* orientational term (all vectors are unit vectors) such that the direction of the vectors is irrelevant. The transition dipole is along the length of the aromatic rings associated with the main dyes (see [665] for more info). This procedure has repeated until we have reached a converged FRET efficiency value for this particular protein conformation, i.e. one where the addition of more dyes no longer changes the mean FRET efficiency by more than some tolerance factor. Once convergence has been reached, the mean value is retained and the process is repeated with the next protein conformation. The conformational convergence (the COCO in COCOFRET) is necessary due to the highly stochastic and conformation-dependent efficiency associated with placing dyes. The result is highly reproducible dye placements in terms of FRET efficiency. If we compare a fixed number of trial placements vs. COCOFRET (see fig. 9.5). To reach convergence can take up to 8000 - 10000 unique dye-dye combinations, although the degree of tolerance is controllable.

There are many adjustable parameters as part of the COCOFRET algorithm, and the entire approach represents a substantial codebase. In the interest of brevity we will not delve more deeply into the underlying implementation details, but from the user perspective the only thing that must be done is to define the two residues where dyes will be placed, which dyes are going where, and from there the remainder of the procedure is entirely automated.

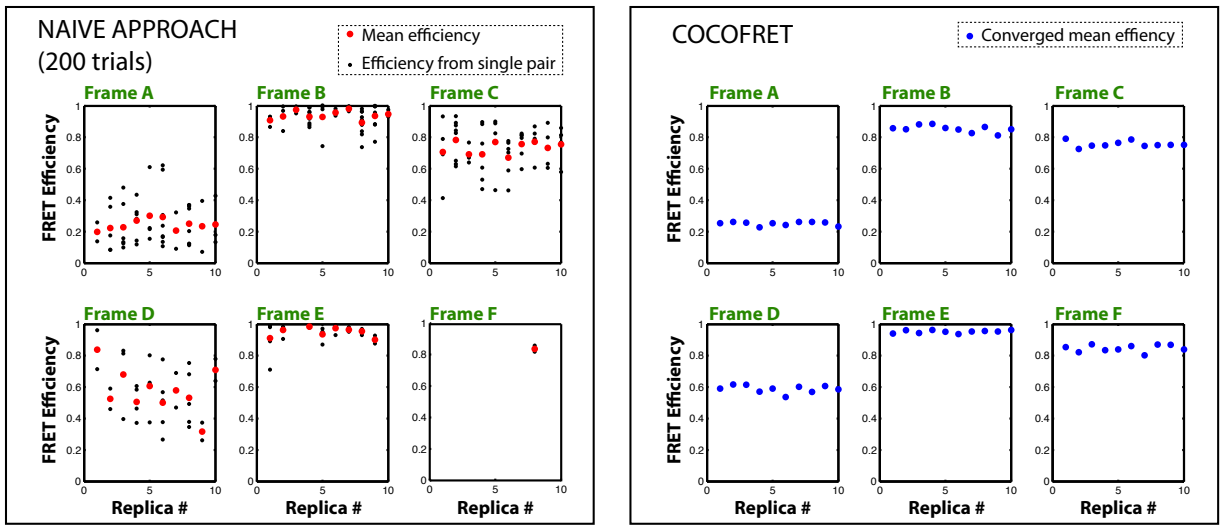

Figure 9.5: COCOFRET reliably and reproducible out-performs a naive dye-placement approach. Note that for intermediate FRET efficiencies especially there is a large stochastic component associated with dye placement, which the conformational convergence helps to remove.

9.3 Discussion

CTraj has been in interactive and continuous development for over two years. As well as the pre-defined analysis routines, CTraj provides the user with direct access to the complete set of atomic coordinates for each frame. These coordinates are accessible via a standardized representation language, as well as via conventional matrix indices. This is a major advantage, as it facilitates rapid development and deployment of novel analysis algorithms in an interactive and well-supported programming environment. The vast majority of the simulation analysis in this thesis has been done using CTraj. We have also used CTraj as the

back-end for a new command-line analysis tool (Robin) which is in development. Taken together, CTraj provides a programmatic, intuitive, and high-performance framework through which all atom simulations of disordered proteins can be readily analysed.

Chapter 10

Future directions I: IDPs

To conclude this section, we will muse on some possible future directions.

10.1 Evolution of IDPs

A challenge associated with IDPs comes from a need to develop entirely new methods for thinking about sequence conservation [70, 182]. Various approaches have been developed to explore evolution in IDPs, but these typically rely on either conventional ideas surrounding conservation or deep, detailed analyses into a single system [249,589]. It is typically challenging to align disordered domains using conventional protein-sequence alignment approaches. This should not be surprising - there is a fundamentally different sequence-function relationship in IDPs than there is in folded proteins. Using alignment approaches (and by extension, evolutionary analyses) that were developed for folded proteins is fundamentally inappropriate. This is equivalent to determining that a monolingual Frankophone is stupid because they can't understand English; we are simply using the wrong language to make the assessment! A new framework must consider the mapping between sequence and ensemble, and

use that ensemble as the feature of conservation, not the underlying amino acid sequence. This is a challenge; it requires accurate ensemble prediction from sequence alone, and while the work in the preceding chapters suggests we are on the right track, more work is required, both in terms of ensemble accuracy and simulation throughput. Where clear features can be identified a simpler solution is to identify conservation of general sequence features, as has been done elegantly by Zarin *et al.* [664]. The challenges with such an approach is it fails to allow for large changes in sequence that have a minimal impact on function. For example, local compaction could be driven by an enrichment in hydrophobic residues, polar residues, or charge interactions - seemingly divergent sequence features with similar conformational outcomes.

Folded and disordered domains have co-evolved, and are typically found within the same protein. Is there a way for us to take advantage of this? Evolution doesn't 'care' about mechanism, it cares about fitness, and if the sequence solution arrived at is not deleterious then it is just as likely as any other possible sequence solution. We propose that an attractive avenue for assessing conservation in IDPs emerges from considering the coupling between folded domains and disordered regions. A protein's impact on fitness cannot be carved up into distinct and discrete modules, but is imparted as an emergent property of the entire protein in the cellular context.

As an example, say we have a protein with a highly conserved folded domain and a disordered C-terminal tail. If the folded domain is genetically deleted cells are non-viable. If the disordered tail is deleted, true fitness is not impacted, and we know this unambiguously (an impossibility, but for the sake of the discussion let's say this is true). A reasonable conclusion one might draw is that this disordered domain is not under selective pressure - it is not necessary for function, has no impact when deleted, and so should be free of selective

pressure. Now imagine we mutate all the residues in this C-terminal tail to aspartic acid - this leads to cell death. We now have a paradox. Mutations in a region that is apparently under no selective pressure will directly impact cellular fitness. How can this be? The answer is simple; the C-terminal tail is under selective pressure to be inert within the cellular context. The sequence space associated with inert-ness may be huge. We may be able to entirely scramble the sequence a million different ways, make drastic deletions or insertions, replace the C-terminal tail with a folded domain, and all of these changes are happily accepted by the cell, but this does not mean the tail is not under selective pressure, it just means that the changes we are making are within the manifold of that selective pressure. What does this mean for protein evolution? We *must* consider a disordered region connected to a highly conserved folded domain to be equally well conserved in terms of function and phenotype, regardless of the apparent degree of conservation.

As a result, we could look for high degrees of conservation in folded domains as a marker for highly conserved disordered regions, regardless of their *actual* amino-acid conservation. Effectively, we treat the protein as being uniformly conserved according to a value that can be determined based on the folded domain, and extrapolated to the disordered region. This provides us with a method to build collections of sequences upon which we have placed *no* constraint on amino acid composition, number of residues, or any other property, yet we believe they have equivalent function. This, in turn, provides us with an enormously powerful collection of data to explore sequence-to-function relationships on a proteomic scale.

We have presented a simplified version of the likely reality. Proteins evolution is under epistatic control [450]. Disordered regions are frequently involved in regulation, and linear motifs have been shown to be gained and lost rapidly across evolution [130, 654]. Consequently, there is the convolving factor that disorder may be a critical route for proteins to

evolve and develop novel functions though changes in regulatory domains, while folded domains remain fixed. This can be thought of as akin to adding new sensors to a robot - the underlying robot's function remains the same, but as sensors (disordered regions) are added, changed, and removed, entirely new functions may emerge.

The reality is likely somewhere between these two extremes. A good starting point would be to identify flexible linkers devoid of interaction motifs. In preliminary work our hypothesis appears to hold water, but significantly further investigation will be needed to assess if this is a reality.

10.2 General Analytical Models for Heteropolymers

As discussed at length at the end of 2 (and will be discussed further in chapter 13), IDPs are not homopolymers. Despite this, the field has achieved enormous leverage from theoretical models developed by Flory, de Gennes, Zimm, Rouse - and many others - that were developed to describe homopolymers [133, 180, 181, 503, 675]. If possible, we should try to collectively evolve beyond these homopolymeric models. Elegant work from Lin & Chan and Swale & Ghosh demonstrates that relevant heteropolymeric behaviour can be captured in theoretical models, albeit for simplified sequences [335, 517]. We have no great insights or advances to share; the inclusion of a Langevin-style stochastic correction term may allow homopolymeric models to be modified to capture a random distribution of residues, but it is entirely possible that the inherent complexity associated with what is effectively a (potentially) 20-parameter analytical model are simply not usable. Nevertheless, a general description of attractive and repulsive regions along a chain seems like a necessary next step for analytical theory.

10.3 Coarse-Grained Models for Heteropolymers

A more tractable approach than the development of novel theory is to design new, coarse grained models that allow for the capture of sequence-specific behaviour while provide throughput that allows thousands of sequences to be analysed a day. While more work is required, in chapter 14 we will discuss our developments on this front.

10.4 Improved Classification of IDPs

Taken together, a general theme that emerges from the preceding three sections is that we may need to revisit our sequence-based classification of IDPs [126, 359, 603]. Amino acid composition is a useful coarse-grained classifier, but a growing body of work suggests the distribution of residues (charged, hydrophobic, glycine, aromatic) has enormous implication for the individual and collective behaviour of IDPs [126, 335–337]. Beyond simply the patterning of individual residues across a sequence, the relative distribution of different sequence motifs is likely critical for function, as described in recent work by Das *et al.* [125]. Taken together, there appears to be a need for a hierarchical framework that considers the patterning and composition of different sequence features across a range of length-scales.

While such an approach would be incredibly powerful, it perhaps glosses over some additional challenges. We have presented a picture of IDPs wherein their amino acid composition and patterning determines conformational behaviour. This is accurate, but ignores three intimately linked but independently confounding axes in the mapping of sequence to ensemble.

The first is **solution conditions**. IDPs are frequently enriched in charged residues, and while we typically imagine the pKa of Asp/Glu to be around 4 and Lysine to be around 11, local charge interactions can (and do) *dramatically* shift pKa values such that local regions can become significantly less charged than they might be naively be expected [225]. Moreover, the cellular milieu is awash with various osmolytes; salts, various phosphate-containing molecules (e.g. NTP and NDP), metabolic intermediates, carbohydrates and a plethora of other compounds [476]. How does the presence of these compounds alter the behaviour of IDPs? Lohman has elegantly shown that simply changing the anion can dramatically influence protein-nucleic acid interactions [344]. How might more significant changes in conditions influence the conformational behaviour of IDPs? More broadly, the impact of crowding on IDPs has been a source of interest for many groups over the last five years - the emerging consensus appears to be that it depends on the IDP and the crowder, which is likely an accurate albeit unhelpful conclusion [385, 556, 579]. In summary, the solution conditions matter immensely. The role of crowding and depletion effects, osmolytes, pH, and temperature remains poorly understand in a general sense, in part because it seems unlikely that there will be a single collection of 'rules' that can describe the relationship between an IDP and it's solution environment.

The second is **binding partners**. Many (indeed, perhaps most) IDPs operate by binding to partners [564, 588]. These partners may folded proteins, nucleic acids, small-molecules, or may themselves be IDPs. A crucial set of questions for IDP functional conservation pertains to the coupled evolution of IDPs with their binding partners. In many cases the binding of an IDP to a folded protein facilitates a coupled folding and binding event, whereby the IDP folds into a well defined structure in the bound state [114, 221, 494, 534]. For IDPs that undergo coupled folding and binding, their evolution is constrained by their behaviour in both the bound and the free state. This may explain why many proteins that undergo coupled

folding and binding fold into a single 20-40 residue α-helix. The structural constraints on the helix are limited, such that there is large set of possible mutations that would still allow for helix formation [26]. Regardless, the interplay of how binding partners may influence sequence constraints in IDPs is of general interest, and we suspect encodes a rich set of information [589].

The third and final challenge is one of **local structure**. We present IDPs as disordered ensembles where, despite well-defined sequence encoded preferences, these unstructured regions behave in a manner akin to flexible polymers. This ignores a basic tenet of protein physics; proteins can fold. Based on our discussion in chapters 1 and 7, folding can be thought of as a cooperative coalescence of unstructured regions driven by a combination of well defined chemistry and local geometry. It is entirely possible that within disordered ensembles local folding events may transiently occur, with local structure (in the conventional sense) flickering into and out of existence on a time-scale faster than is observable by any solution technique. If we accept this postulate, then there may also be single point mutations which, in terms of general sequence properties appear insignificant, but represent a tipping point on the edge of structure formation. This structure could exist in terms of intramolecular interactions, or in terms of intermolecular interactions, the latter providing a potential explanation for why certain disease-associated point mutations can dramatically change a disordered protein's propensity to form amyloid-like fibrils (although we emphasise this does not preclude other mechanisms between mutations and amyloid formation, of which there are likely many) [83, 399, 443, 459]. We and others speculate that the formation of cross-β spine fibrils represents an intrinsic energetic ground state for proteins at high

[26] As an aside, such a model is supportive of a general framework whereby a stably folded protein facilitates evolution by allowing mutational tolerance. Many single point mutations in folded proteins are permissible because their associated hit in folding stability is insufficient to unfold the protein. This provides an answer to the seemingly unimportant but perhaps quite relevant question of why many folded proteins remain stable at 1.5 x the normal growth temperature of the organism.

concentration driven by backbone interactions, and is not necessarily a state that is related to function [26, 55, 296]. The challenge with these tipping-point mutations is that they introduce deep discontinuities into the state space of IDPs - small perturbations to sequence lead to enormous perturbations in conformation. Identifying those discontinuities will likely be challenging, but represent an important caveat when attempting to classify disordered proteins based on sequence alone.

Part III

Collective Chain Behaviour

Chapter 11

Sequence determinants of intracellular phase separation by complex coacervation of a disordered protein

The following section is taken from the paper **Sequence determinants of intracellular phase separation by complex coacervation of a disordered protein** by C.W. Pak, M. Kosno, A.S. Holehouse, S.B. Padrick, A. Mittal, R. Ali, A.A. Yunus, D.R. Liu, R.V. Pappu, M.K. Rosen. This was published in *Molecular Cell*, Vol. 63, pages 72 - 85, in July 2016. The text has been expanded to include additional detail. Author contributions were as follows Conceptualization, C.W.P., R.V.P., and M.K.R.; Methodology, C.W.P., S.B.P., **A.S.H.**, R.V.P., and M.K.R.; Software, C.W.P., S.B.P., A.M., and **A.S.H.**; Formal Analysis, C.W.P. and **A.S.H.**; Investigation, C.W.P., M.K., S.B.P., **A.S.H.**, A.M., and R.A.; Resources, A.A.Y. and D.R.L.; Writing C.W.P., S.B.P., **A.S.H.**, R.V.P., and M.K.R.; Visualization, C.W.P., S.B.P., **A.S.H.**, R.V.P., and M.K.R. .

11.1 Introduction

Membraneless organelles such as nucleoli, Cajal bodies, and stress granules are involved in diverse biological processes, from ribosome assembly, to gene regulation to signal transduction [89, 106, 612, 631, 671]. These micron-sized organelles are found throughout the cell, and like their lipid-membrane-bound counterparts, provide distinct cellular compartments, concentrate select proteins and nucleic acids, and may localize specific biological reactions [27, 251, 558]. We define membraneless organelles as intracellular condensates that encapsulate a set of proteins and RNA and allow the passive diffusion of smaller species into and out of the organelles.

Many membraneless organelles have liquid-like physical properties. They are spherical, undergo fusion/fission, drip under shear stress, and exchange contents rapidly with the surrounding medium [251]. Such structures include P granules in *C. elegans* embryos, nucleoli in *Xenopus* oocytes and *C. elegans* embryos, germ granules, stress granules in mammalian cells, and ribonucleoprotein granules in fungi [41, 65, 66, 162, 172, 307, 421, 625, 633, 668].

Analogous liquid-like structures have also been generated by expressing various multivalent or disordered proteins [330, 399, 420, 421, 443]. Their physical properties and condensation/dissolution behavior suggest that at least some of these structures assemble via a liquid-liquid phase transition [65, 66, 330, 421, 633]

Many membrane-less organelles are enriched in proteins containing large intrinsically disordered regions (IDRs) [158]. These regions lack persistent three-dimensional structure, but often contain multiple weakly adhesive sequences that include hydrophobic/aromatic, uncharged polar, and/or charged residues [608]. IDRs that lack charged residues phase separate due to a preference for homotypic self-associations over interactions with the solvent [67].

In polymer science, phase separation driven by homotypic interactions is known as simple coacervation. This process involves formation of a dense phase enriched in a single polymer that is in equilibrium with a polymer-depleted phase [610]. Most IDRs known to phase separate *in vitro* do so via simple coacervation. Tropoelastin, an extracellular matrix protein composed of imperfect repeats of VGVAPG, undergoes simple coacervation *in vitro* through interactions among its hydrophobic residues [171, 383, 602]. IDRs that phase separate are also enriched in RNA binding proteins, where they often possess a limited set of amino acids - F, Y, S, G, Q, N [113, 217, 282]. Ddx4, an RNA binding protein localized to germ granules, contains IDRs that phase separate *in vitro* and in cells. In Ddx4, electrostatic interactions between clusters of opposing charge and cation-pi interactions between FG and RG motifs have been invoked as important driving forces for phase separation [421]. Two RNA binding proteins, FUS and hnRNPA1, also phase separate *in vitro* and in cells, and Tyr residues promote recruitment of FUS into RNA granules [79, 217, 282, 399, 443]. Mutation of multiple Tyr to Ser in the mitotic spindle protein BugZ reduces its ability to phase separate [266]. At high concentrations, many proteins containing low complexity IDRs, including FUS and hnRNPA1/2, form amyloid-like filaments consisting of repeated cross-beta strand elements [217, 282, 338, 399]. An intriguing postulate is that cross-beta elements, when occurring in small numbers between otherwise disordered chains, constitute the adhesive structures that promote liquid phase separation (i.e. are the structural correlates of the thermodynamic driving forces mentioned above) [659]. Such a hypothesis is unable to explain how PR_{20} is able to phase separate, given the proline-rich polymers inability to form β-sheets, although the anion dependence hints at a cation-chelation effect [51]. In proteins containing both modular binding domains and IDRs, module-ligand and IDR-IDR interactions act cooperatively to promote phase separation [338].

When polymers are enriched in one type of charge, they repel one another, and this inhibits simple coacervation. However, when mixed with multivalent counterions, highly charged polymers phase separate via a process known as complex coacervation. The oppositely charged molecules phase separate together, forming an electroneutral polymer-rich droplet phase where both species are highly concentrated, and this dense phase is in equilibrium with a polymer-depleted bulk phase [610].

In this work we have examined the intracellular phase separation of a negatively charged IDR, and find that it forms nuclear bodies via complex coacervation. Our investigations are based on the serendipitous observation that the disordered intracellular domain of the adhesion receptor Nephrin forms nuclear bodies when expressed as an autonomous entity in mammalian cells. Using quantitative microscopy, we show that nuclear bodies formed by the Nephrin intracellular domain (NICD) behave as phase separated liquid droplets, and are likely novel nuclear structures. Cellular and biochemical data indicate that bodies/droplets form through non-specific associations of the negatively charged NICD with positively charged partners. Using deletion mutants and *de novo* sequence designs, we show that NICD phase separation is promoted by one or more blocks of high negative charge density and by aromatic/hydrophobic residues distributed along the sequence. The data demonstrate that the amino acid composition of NICD is more important than the precise sequence for phase separation. Using bioinformatics analysis, we identified 443 unique IDRs within the human proteome with sequence features that are similar to that of NICD. This suggests that complex coacervation of negatively charged IDRs may contribute generally to the formation of membrane-less organelles and clusters of membrane receptors.

11.2 Methods

11.2.1 Analysis of NICD Nuclear Bodies in Cells

Human NICD was PCR amplified from a plasmid encoding CD16/7-Nephrin-GFP (a kind gift from Drs. Jones and Pawson) and inserted into a CMV promoter driven EGFP fusion vector. NICD mutants were produced using PCR with appropriate primers, and confirmed by sequencing. HeLa cells were cultured in DMEM supplemented with 10% FBS, GlutaMAX and Pen/Strep mix. NEAT1$^{-/-}$ and NEAT1 $^{+/+}$ MEFs were cultured in a 1:1 mixture of DMEM and Ham's F-12, with the same additives. For imaging experiments, cells were cultured in glass bottom dishes prior to transfection using Lipofectamine 2000 (Thermo Fisher). Live HeLa cells were imaged using a Zeiss LSM510 confocal microscope or Deltavision RT widefield microscope on a heated stage. For FRAP analysis, nuclear bodies were photobleached using a 488 nm laser, and custom ImageJ and FIJI scripts were used to determine the area of and normalized intensity within the photobleached region at each time point. NICD nuclear bodies in fixed HeLa cells were imaged in 3D at high-resolution using spinning disc confocal microscopy, and analysed in a semi-automated fashion. All images were background-subtracted and flatfield-corrected, and cellular autofluorescence was determined to be negligible. The volumes of nuclear bodies (V_{NBs}) larger than the transverse resolution limit were determined from their diameter and an assumption of spherical shape. Volumes of smaller nuclear bodies and intensities of all nuclear bodies (INBs) were calculated using a calibration that relates the effects of the point spread function on intensity. Molar fraction was calculated as $([V_{total,NBs}] \times [I_{mean,NBs}]) / ([V_{total,nucleus}] \times [I_{mean,nucleoplasm}])$, and partition coefficients for each nuclear body were calculated as $(I_{mean,NBs}) / (I_{mean,nucleoplasm})$.

Cell populations were scored for the presence of NICD nuclear bodies in transfected HeLa cells (which are fluorescent), when imaged with a Deltavision RT widefield microscope. Quantification of nuclear body formation was performed at unmatched and matched expression levels; data are shown as mean ± SEM. We limited this analysis to nuclear bodies larger than the transverse point spread function of our microscope and took care to correct for the size-dependent decrease in apparent intensity due to the point spread function (fig. 11.1).

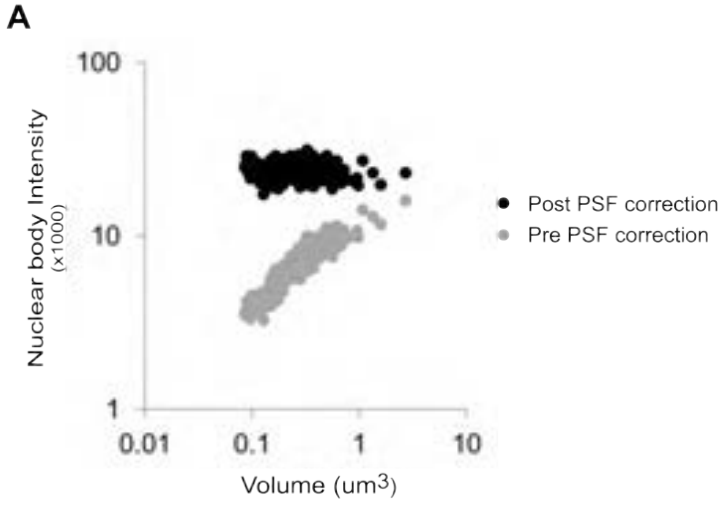

Figure 11.1: Quantification of nuclear body intensity ($n = 239$ nuclear bodies from 30 cells) before (gray) and after (black) correcting for the effect of the point spread function. Before correcting for the point spread function, nuclear body intensity scales with volume. After correcting for the point spread function, nuclear body intensity is independent of volume.

11.2.2 Atomistic Simulations of NICD

All simulations were performed using the CAMPARI Monte Carlo (MC) modeling suite (http://campari.sourceforge.net), which uses the ABSINTH implicit solvent model and force

field paradigm [613]. Each simulation was initiated from a random, non-overlapping starting conformation. Single protein simulations used temperature replica exchange Monte Carlo (T-REMC), and all analysis was performed on conformations from the 310 K ensemble. Dimer simulations were run at 310 K. For further discussion on the ABSINTH forcefield please see chapter 2.

11.2.3 Sequence Analysis

Sequence analysis for identifying CIEs was performed using localCIDER (see chapter 4). The parameters used to define CIEs were based on previous work in polymer and polyampholyte physics [126]. Contributions of specific residue types to nuclear body formation were determined by calculating Pearson's correlation coefficients, which assessed how well the loss of unique combinations of residues correlated with changes in the formation of nuclear bodies. All unique combinations of residues were considered. Variance in the experimental data was accounted for by calculating a distribution of correlation values using sub-sampled datasets. Significant differences were evaluated using unpaired Student's t-tests, and residues which were consistently enriched relative to a normalized background were identified.

11.2.4 Protein Production

Recombinant NICD protein was produced from a codon-optimized human wild type NICD sequence with a C-terminal Tev-cleavable His8 tag, expressed from a modified pMAL-C2 vector with two N-terminal tandem Tev-cleavable MBP domains. Mutant NICD sequences (charge: CC, CS, CB_C) were subcloned from mammalian expression vectors into the same expression system, and confirmed by sequencing. NICD and scGFPs proteins were expressed

in E. coli BL21(DE3)T1^{R} and purified using a combination of affinity, ion exchange and gel filtration chromatography. NICD was labelled with Alexa Fluor 568 on the single endogenous cysteine using maleimide chemistry. Two synthetic peptides, CR7 (sequence = CRRRRRRR), and CR20 (sequence = CRRRRRRRRRRRRRRRRRRR), were synthesized and purified by GenScript (NJ, USA).

11.2.5 *In vitro* NICD Phase Behaviour

In vitro phase separation assays were performed by mixing NICD and/or scGFP directly in a 384-well plate to give the indicated concentrations, incubating for 24 hours at room temperature in the dark, and imaging using fluorescence confocal microscopy. Saturation concentrations were determined in one of two approaches. In an imaging based approach, a grid of NICD and scGFP concentrations was prepared, imaged, and automatically scored for the presence of phase separated droplets using custom ImageJ, FIJI and Python scripts. Images with obvious inclusions were manually rejected. In an alternative approach, solutions of Alexa-568 NICD/scGFPs were clarified by centrifugation and the residual concentration of NICD/scGFP in the supernatant was determined by fluorescence using standards of known protein concentrations. Phase separated droplet sizes and normalized intensities were measured using custom ImageJ scripts, where droplet intensity values were normalized to those of WT NICD + scGFP(+15). Data are represented as mean ± SEM. FRAP analysis of phase separated droplets of NICD/scGFP used a 405 nm laser for photobleaching and custom ImageJ scripts to determine the area of and normalized intensity within the photobleached region at each time point.

11.3 Results

11.3.1 NICD Forms Nuclear Bodies that are Phase-Separated Liquids

The adaptor protein, Nck, and the actin nucleation-promoting factor, N-WASP, can phase separate when mixed together, due to multivalent interactions between SH3 domains on Nck and proline rich motifs on N-WASP [330]. The Nck SH2 domain can also bind to phosphotyrosine sites on phosphorylated NICD. *In vitro* this interaction facilitates phase separation by scaffolding multiple Nck proteins, thereby increasing the effective valency of Nck SH3 domains [29]. We initially sought to recapitulate these phenomena in HeLa cells by overexpressing the NICD (residues 1077-1241) tagged at its C-terminus with a fluorescent protein (YPet) (fig. 11.2A, red box). We expressed either wild type NICD or, as an intended control, a non-phosphorylatable mutant (Y3F). Surprisingly, both proteins formed micron-scale nuclear bodies (figs. 11.2B and C; note that percent cells showing puncta is normalized to that of WT NICD throughout this report). Thus, these structures do not require the binding of Nck to phosphotyrosine motifs on Nephrin, and are unrelated to the previously described phase separation of the ternary system.

NICD nuclear bodies varied in number across cells, varied in size within a given cell, and had liquid-like physical properties. They were spherical at the resolution of light microscopy (fig 11.2B). NICD-YPet fluorescence in bodies recovered rapidly after photobleaching, with $\tau < 1s$ for structures $\sim$1 µm in diameter, consistent with a molecular diffusion constant of $0.5\,\mu m^2\,s^{-1}$ (fig. 11.2D) and rapid exchange with the surrounding nucleoplasm. NICD-YPet fluorescence returned to only 90% of its initial value due to photobleaching of the cells during

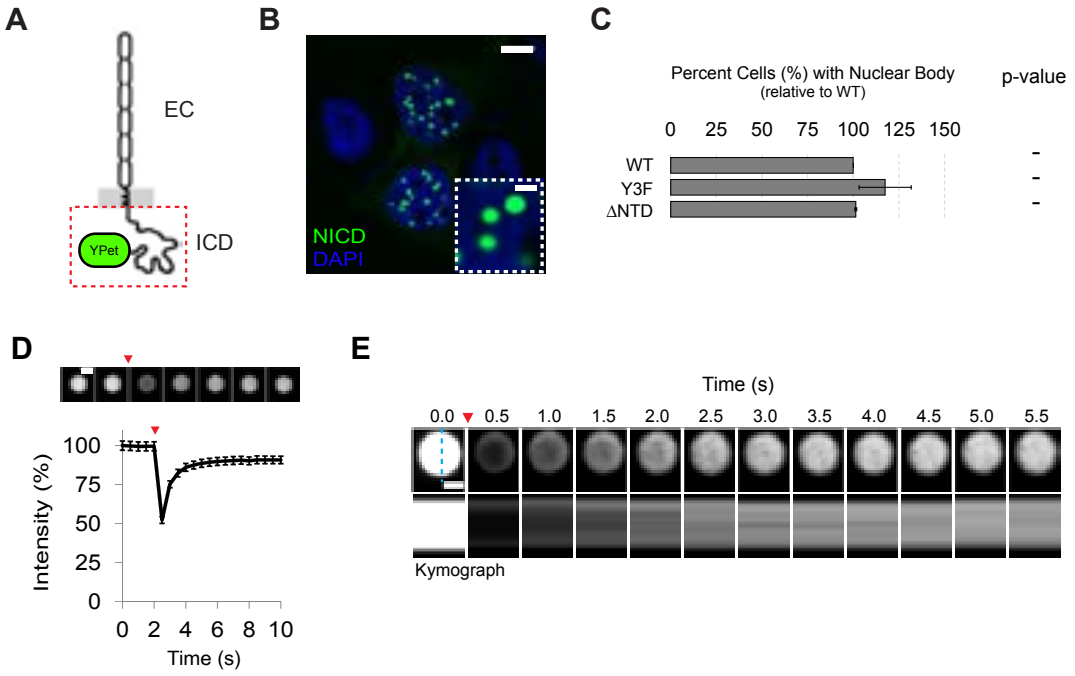

Figure 11.2: NICD nuclear bodies are phase separated liquids. (A) Schematic representation of nephrin, including extracellular region (EC) and intracellular cytoplasmic domain (ICD). NICD (red box), expressed as a soluble protein, is C-terminally tagged with YPet. (B) Spherical micron-sized nuclear bodies (inset) form in the nuclei of HeLa cells expressing NICD (green). Nuclei were stained with DAPI (blue). Scale bars in main and inset panels are 5 and 1 µm, respectively. (C) Normalized (to WT) percent of HeLa cells containing nuclear bodies when expressing NICD mutants Y3F (Y100F, Y117F, and Y141F) or ΔNTD (residues 63-166). Data are represented as mean ± SEM (p values for comparison to WT NICD: '-' indicates $p > 0.05$). (D) Confocal imaging and quantification of NICD fluorescence recovery after photobleaching ($n = 34$ bodies). Red arrowhead marks photobleaching event. Data are represented as mean ± SEM. Scale bar = 1 µm. (E) Upper: recovery of fluorescence intensity in a larger nuclear body. Red arrowhead marks photobleaching event. Scale bar = 1 µm. Lower: kymographs of fluorescence recovery across the diameter of the nuclear body (cyan dashed line).

the recovery period (equivalent time lapse imaging decreased the intensity of unperturbed nuclear bodies by ∼14 %, not shown). In larger bodies monitored by confocal microscopy, fluorescence recovery was observed initially at the periphery, followed by radial spreading to the center (fig. 11.2E). Thus, these bodies are filled objects, and constituent NICD molecules exchange with the nucleoplasm faster than they diffuse within the bodies. Finally, NICD nuclear bodies also fuse rapidly ($<$ 10 s, fig. 11.3F, left panel), and with conservation of volume (fig. 11.3F, right panel). Thus, NICD nuclear bodies behave as liquid droplets.

To examine the concentration dependence of body formation, we initially imaged transiently transfected cells over time as they began to express NICD. In most cells, numerous, small nuclear bodies formed suddenly ∼6-8 h after transfection, when expression levels were relatively low (fig. 11.3G; the time point preceding the first instance of a nuclear body was set to $t = 0$). Nuclear bodies increased in size and decreased in number over time (fig. 11.3G), likely due in part to merging (cf. fig. 11.3F). We also examined how the total summed volume of NICD nuclear bodies in a given cell varied with nuclear expression level. HeLa cells expressing different levels of NICD were imaged in 3D by confocal microscopy (fig. 11.3H, left panels (a-d)). For each cell, we determined the total fraction of the nucleus occupied by NICD bodies. Below an average nuclear fluorescence intensity of ∼600-750 A.U., cells never contained observable nuclear bodies (fig. 11.3H, marker a). However, bodies appeared sharply above this value (fig. 11.3 panels H and I, marker b), and total nuclear volume fraction increased steadily with increasing nuclear fluorescence (NICD concentration; (fig. 11.3 panels H and I, markers c and d)). Nuclear bodies comprised 1-3.5% of the nuclear volume (fig 11.3I), and retained up to ∼25-40% of total nuclear NICD molecules (fig 11.3J). We never observed bodies comprising $> 3.5\%$ of the nuclear volume likely due to toxicity of NICD at extremely high concentrations. These data demonstrate that NICD phase separates to generate liquid-like nuclear bodies.

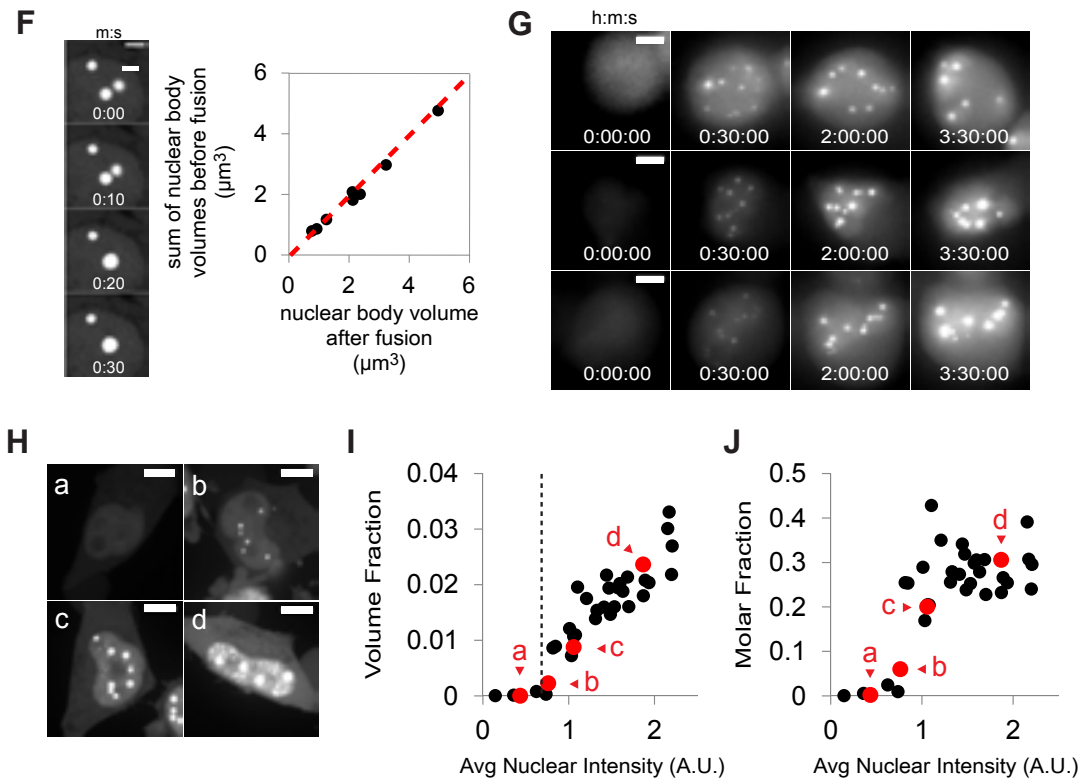

Figure 11.3: (F) Left: time lapse imaging showing fusion of two nuclear bodies. Scale bar = 2 µm. Right: volume is conserved during coalescence ($n = 8$ bodies) (slope of red dashed line = 1). (G) Time lapse imaging of nuclear body assembly in transiently transfected HeLa cells expressing NICD. Scale bar = 5 µm. (H) Maximum projection images of HeLa cells expressing different levels of NICD (expression level increases from (a) to (d)). Scale bar = 5 µm. (I, J) Quantification of volume fraction (I) and molar fraction (J) of nuclear bodies in cells ($n = 36$) expressing different levels of NICD. Each symbol represents a single cell. Volume and molar fraction of cells shown in (a) through (d) are indicated by red symbols. Saturation concentration at ∼600-750 arbitrary units is indicated by the black vertical dashed line in I.

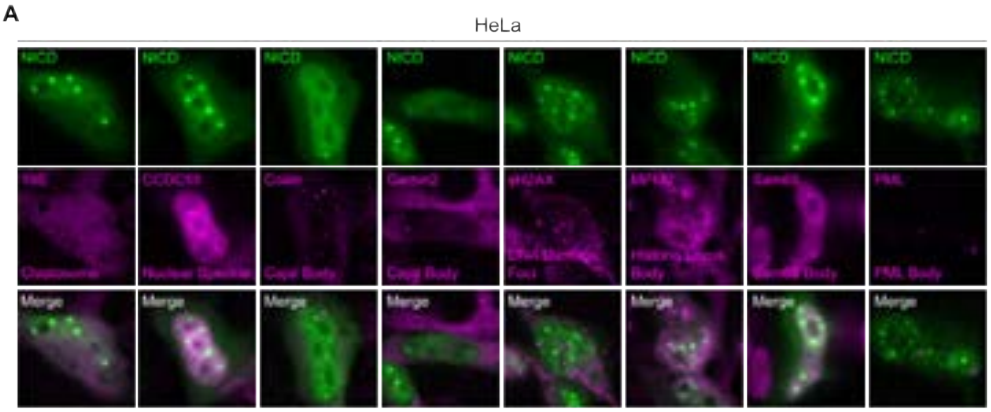

Figure 11.4: (A) HeLa cells expressing NICD-YPet (green) were stained with antibodies (magenta) to visualize marker proteins (in parenthesis) characteristic of the indicated endogenous nuclear bodies: clastosome (19S), nuclear speckles (CCDC55), Cajal bodies (coilin/gemin2), DNA damage foci (γH2AX), histone locus bodies (MPM2), Sam68 bodies (Sam68), PML bodies (PML), and paraspeckles (NONO/SFPQ; shown in 11.5B))

We next asked if NICD nuclear bodies are formed *de novo* or if NICD is absorbed into an existing nuclear body. In HeLa cells, NICD bodies do not co-localize with markers of Cajal bodies, nucleoli, PML bodies or several other nuclear puncta (fig. 11.4 and fig. 11.5). However, they co-localize with nuclear paraspeckle markers NONO/p54nrb and SFPQ/PSF (fig. 11.5B). Nevertheless, NICD bodies form equally well in cells where NEAT1 has been deleted (fig. 11.5C), which lack paraspeckles (fig. 11.5D and Clemson *et al.* [109]), and in parental cells containing NEAT1. Thus, they are probably *de novo* structures that absorb NONO/p54nrb and SFPQ/PSF (and perhaps paraspeckles). We note that the sharp appearance of NICD nuclear bodies with expression level is also inconsistent with simple partitioning into pre-existing paraspeckles. Although NICD nuclear bodies co-localize with paraspeckle proteins, they are distinct from classical NEAT1-dependent paraspeckles.

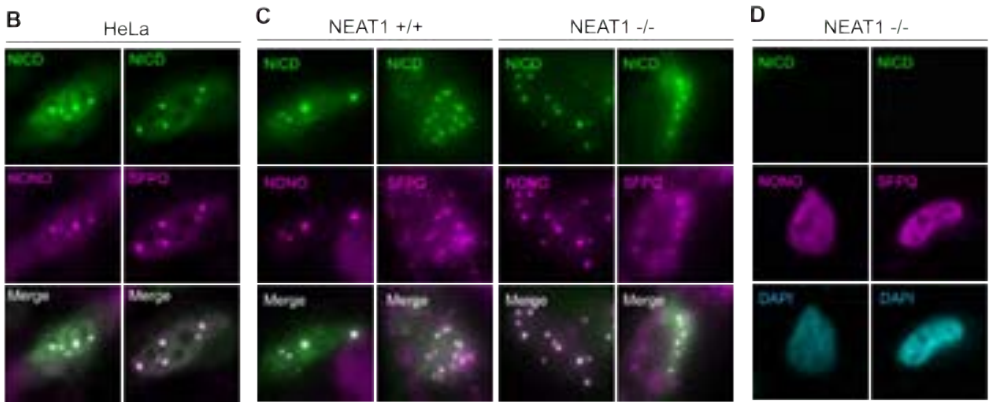

Figure 11.5: (B) NICD (green) nuclear bodies co-localized with known paraspeckle proteins, NONO and SFPQ (magenta), in HeLa cells. (C) NICD (green) nuclear bodies form in NEAT1 -/- mouse embryonic fibroblasts (MEFs), which lack paraspeckles, and in parental NEAT1 +/+ MEFs. NICD co-localizes with NONO and SFPQ (magenta) in NEAT1 -/- and +/+ cells. (D) NONO and SFPQ (magenta) are diffuse in nuclei of NEAT1 -/- MEFs that have not formed NICD nuclear bodies.

11.3.2 NICD Nuclear Bodies Form According to Complex Coacervation

In simple coacervation, when the concentration of a solute is increased to its solubility limit, the solution separates into two phases: a solute-rich phase (droplet phase) of smaller volume, and a solute-poor phase (bulk phase) of larger volume. As the total solute is increased beyond the solubility limit, the volume of the droplet phase increases at the expense of the bulk phase. Importantly, the concentrations of both phases remain constant, so that the partition coefficient of the solute ([solute]droplet/[solute]bulk) is invariant to the total amount of solute in the system [140]. We examined the NICD-YPet fluorescence intensity within nuclear bodies and the surrounding nucleoplasm in a population of cells expressing different levels

of NICD. NICD body fluorescence intensity is relatively independent of size (fig. 11.1), and essentially invariant with total nuclear fluorescence (fig 11.1A). In contrast, fluorescence intensity in the surrounding nucleoplasm is proportional to nuclear NICD expression level (fig. 11.6B). Thus, the partition coefficient of NICD varies with expression level (fig. 11.6C), a behaviour that is inconsistent with simple coacervation.

NICD is negatively charged, with an estimated isoelectric point (pI) of 4.3, net charge of -21 and net charge per residue of -0.13 at neutral pH, and fraction of charged residues (FCR) of 0.30. Initial atomistic simulations of NICD in the absence of multivalent counterions revealed the presence of intermolecular repulsions that were quantifiable in terms of the large distances between pairs of NICD molecules. We hypothesized that NICD phase separation might require binding to cellular targets, akin to complex coacervation. We tested this postulate by performing a series of simulations of NICD with oligo-lysine and oligo-arginine peptides of different valencies and observed multivalent counterion-mediated self-association of NICD, demonstrating that NICD molecules can only self-associate if electrostatic repulsions are diminished via counterion-mediated charge neutralization (fig. 11.6 and 11.22). This neutralization can occur through the accumulation of positively charged protein/peptide counterions along NICD. Such a mechanism is consistent with complex coacervation [292].

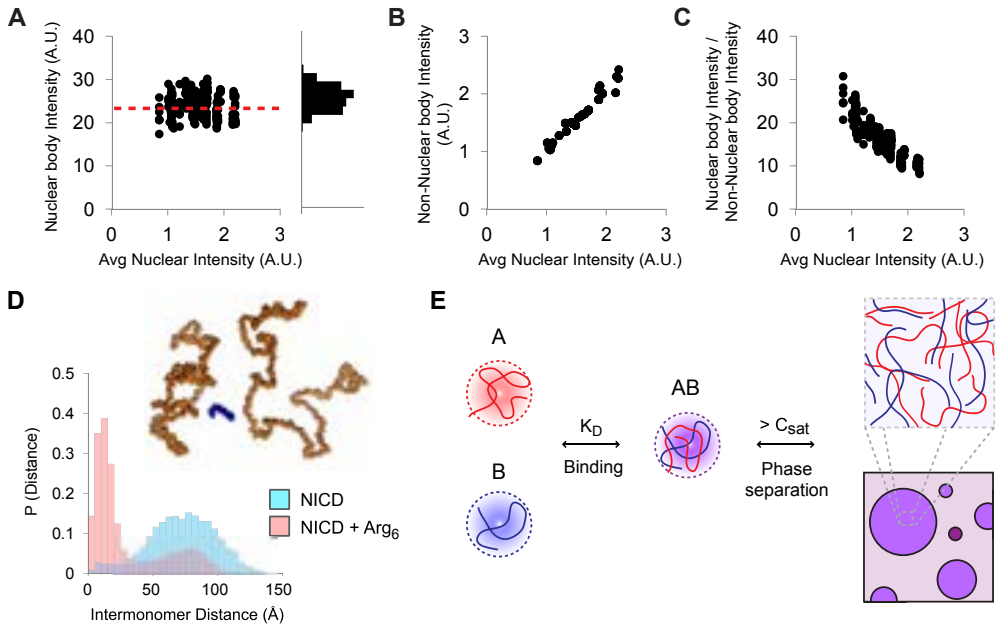

Figure 11.6: NICD nuclear bodies form according to complex coacervation. (A) Left: nuclear body intensity in HeLa cells expressing different levels of NICD. Each symbol represents an individual nuclear body. Dashed red line indicates average for 239 nuclear bodies from 30 cells. Right: histogram of nuclear body intensities.(B) NICD intensity in the surrounding nucleoplasm (non-nuclear body intensity) for 30 cells, each indicated by a single symbol. (C) NICD partition coefficient (nuclear body intensity / non-nuclear body intensity) from cells expressing different levels of NICD (239 nuclear bodies from 30 cells). (D) Histograms of the distances of closest approach between NICD molecules with and without the inclusion of Arg_6 peptides in atomistic Monte Carlo simulations. Purple bars indicate overlap between the two histograms. Inset: representative snapshot of the association of pairs of NICD molecules mediated by Arg_6 peptide (see also fig. 11.22). (E) Complex coacervation model of NICD nuclear body formation. Species A (red) binds to a partner B (blue) with an affinity (K_D). The AB complex phase separates at concentrations greater than or equal to C_{sat}, to form droplets enriched in A and B.

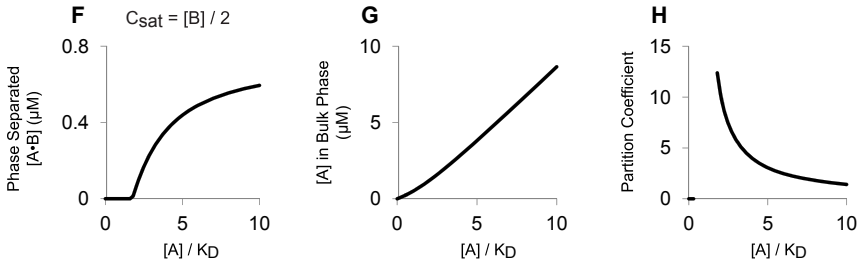

Figure 11.7: NICD nuclear bodies form according to complex coacervation. (F) Modeling complex coacervation of AB. When $K_D = C_{\text{sat}} = \dfrac{[B]}{2}$, phase separated AB appears sharply above a threshold concentration of A. At higher concentrations of A, the amount of phase separated AB plateaus. (G) Except at low concentrations of total A, the concentration of unbound A, which remains in the bulk phase, rises nearly linearly with total amount of A added. (H) When the droplet phase is of constant concentration (cf. fig 11.6A), the partition coefficient of A (as AB complex) decreases with total A.

We modelled complex coacervation by considering two species, A and B, that are soluble individually, but dimerize according to an equilibrium dissociation constant, K_D, to generate a complex, AB, that phase separate beyond a saturation concentration, C_{sat} (fig. 11.6E). As shown in fig. 11.7F, when K_D is similar to C_{sat} and B is limiting, the addition of A does not produce phase separated AB initially. However, when a threshold concentration of A is reached, AB phase separates to an extent that rises with additional A. At very high concentrations of A, the amount of phase separated AB plateaus as the concentration of B becomes limiting. Throughout this process, the concentration of unbound A rises nearly linearly with the total concentration of A (fig. 11.7G). Assuming the concentration of AB in the droplet phase is constant, this increase in the concentration of unbound A results in a steady decrease in the relative partitioning of A (as species AB) into the droplet phase (fig. 11.7H). These behaviours - a sharp appearance of the droplet phase with increasing amounts

of A, followed by the increasing droplet volume with additional A; a steady increase in the level of A in the bulk phase as the total concentration of A increases; and the consequent decrease in partition coefficient of A as the total concentration of A rises - are identical to our experimental observations in fig. 11.3I, 11.6B and 11.6C, respectively, and are observed qualitatively over a range of parameter values. Taken together, these results suggest that NICD may form nuclear bodies via complex coacervation.

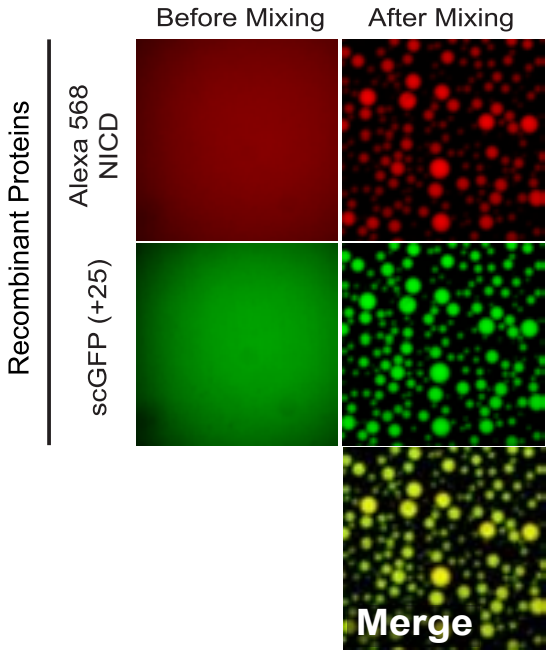

Figure 11.8: Alexa 568-NICD and scGFP(+25) form homogeneous solutions individually, but phase separate together into micron-sized spherical droplets when mixed.

To test this hypothesis *in vitro*, we recombinantly expressed and purified human NICD. Ultraviolet circular dichroism, nuclear magnetic resonance spectroscopy, and gel filtration data indicate that the protein is disordered and likely monomeric in solution. Additionally, atomistic simulations show that NICD adopts conformational ensembles that are expanded

relative to globular, folded domains (fig. 11.10D). Isolated Alexa-568 labelled NICD remained uniformly dispersed in solution with no evidence of phase separation by fluorescence or light microscopy under a range of pH, salt and buffer conditions, both at room temperature and at 4°C (fig 11.8, left panels; see also fig. 11.10E). However, solutions containing NICD plus a supercharged GFP mutant having net surface charge of +25 [scGFP(+25)], as a generic positively-charged macro-ion, were opalescent, and contained spherical droplets enriched in both proteins (fig. 11.8, right panels; see also fig. 11.10E, rightmost panels) [375]. The inverse capillary velocity, which describes the ratio of effective viscosity to surface tension, was estimated to be $0.25\,\mathrm{s}\,\mu\mathrm{m}^{-1}$ for phase separated droplets of wild type NICD and scGFP(+20), a value similar to that of phase separated droplets of the disordered protein, LAF-1 [162]. NICD also phase separated with positively charged oligoArg peptides. Thus, *in vitro*, NICD phase separates via complex coacervation.

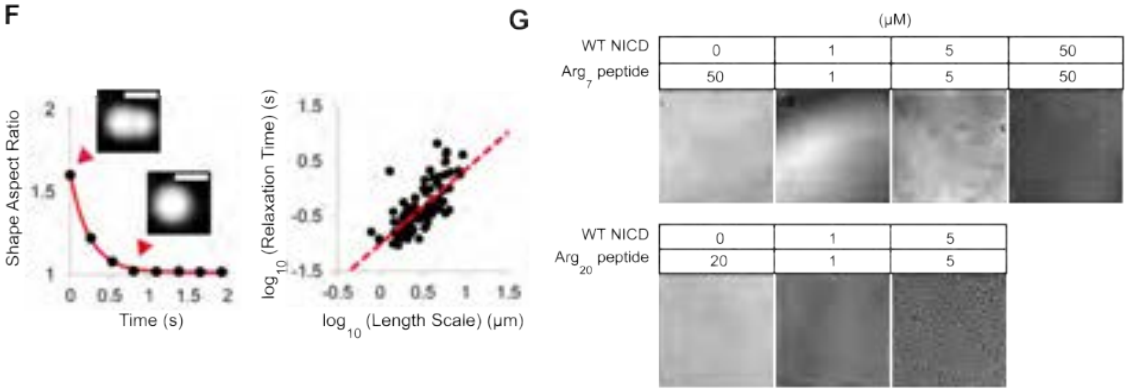

Figure 11.9: (F) Left: representative example of the time course of fusion of phase separated droplets *in vitro* (WT NICD with scGFP(+20)). Right: time constants of relaxation are plotted versus length scale of droplets. The inverse capillary velocity is $0.25\,\mathrm{s\,\mu m^{-1}}$. (G) Solutions containing equimolar concentrations of NICD (1, 5, and 50 µM) and Arg_7 peptide (1, 5, and 50 µM) or Arg_{20} peptide (1 and 5 µM) were imaged for evidence of phase separation by light microscopy. Both peptides were homogeneously distributed in isolation (50 and 20 µM for Arg_7 and Arg_{20}, respectively), but phase separated with NICD. In both cases, higher concentrations of the two species resulted in more robust phase separation (rightmost panel).

11.3.3 Phase Separation of NICD is Promoted by Positive Charge in Partners

Theories and experimental studies with non-biological polymers indicate that in complex coacervation the driving force for phase separation and properties of the resulting concentrated phase depend on the charge density and charge valence of the complexing counterions [610]. We examined these issues using a larger panel of supercharged GFPs: -7 (WT), +7, +9, +15, +20, +25, +36.

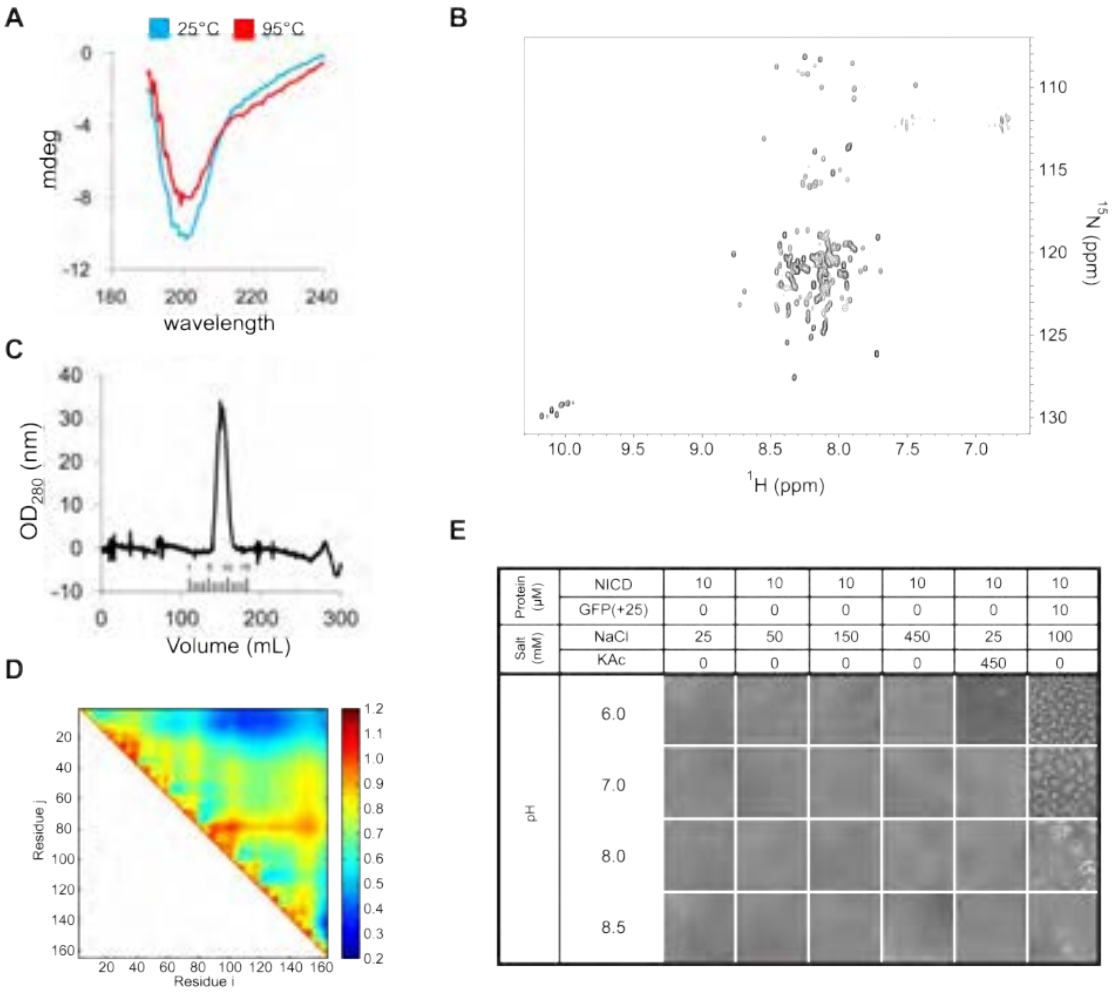

Figure 11.10: (A) CD spectra of NICD at 25°C (cyan) and 95°C (red) shows minor differences indicating low secondary structure content. (B) 1H/15N HSQC spectrum of 87 µM ^{15}N-labelled NICD recorded at 25°C on an 800 MHz NMR spectrometer. The poor chemical shift dispersion suggests a lack of persistent secondary structure. (C) NICD migrates as a single monodisperse peak on a Superdex gel filtration column (D) Scaling maps are consistent with a disordered structure. (E) Isolated recombinant NICD remains soluble under a wide range of conditions, as imaged by light microscopy. In 150 mM NaCl and at pH 6.0 - 8.5, phase separated droplets were observed when supercharged GFP (+25) was added to NICD (rightmost column).

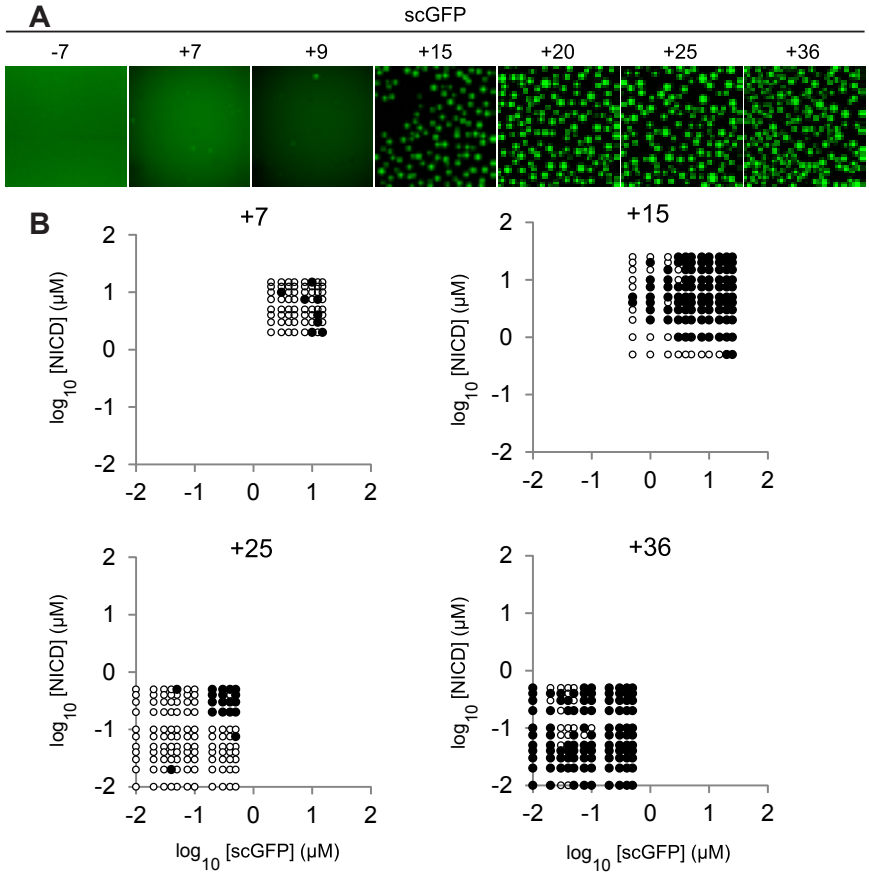

Figure 11.11: Phase separation of NICD is promoted by positive charge in partners. (A) Solutions containing 5 µM wild-type GFP (-7) or supercharged GFPs (scGFPs: +7, +9, +15, +20, +25, +36) and 5 µM NICD were imaged by fluorescence microscopy. (B) Different concentrations of NICD and scGFPs were mixed and scored for phase separation (black circle: phase separated; white circle: not phase separated).

We first examined the charge dependence of the saturation concentration for phase separation using two different measures - the appearance of phase separated droplets in fluorescence images across a grid of scGFP and NICD concentrations (fig 11.11A and 11.11B), and the residual concentration of scGFP in bulk solution after phase separation of 5 µM each of NICD and scGFP (fig 11.12C). The two measures showed good agreement. Both revealed that the

saturation concentration decreases with increasing positive charge on scGFP, from ∼5 µM for scGFP(+7) to 0.01-0.1 µM for scGFP(+36). WT GFP did not promote phase separation up to 10 µM concentrations. The concentration of NICD and (+)-scGFPs in droplets remained constant across the series (fig. 11.12D). Therefore, the partition coefficient also increased with scGFP charge (fig. 11.12E). *In vitro* C_{in}/C_{out} ratios (550 - 10,000) are significantly higher than the maximum *in vivo* C_{in}/C_{out} ratio (30), perhaps due to differences in solubility of the proteins in the cellular environment versus aqueous buffer or to a lower average positive charge on cellular ligands. Finally, FRAP analysis revealed that the movement of scGFP within droplets and its exchange between the droplet and bulk phase decreased with increasing positive charge (fig. 11.12F), suggesting increased strength of intermolecular interactions with higher charge. Thus, the charge of the added species affects the saturation concentration, degree of partitioning, and dynamics of complex coacervates formed by NICD.

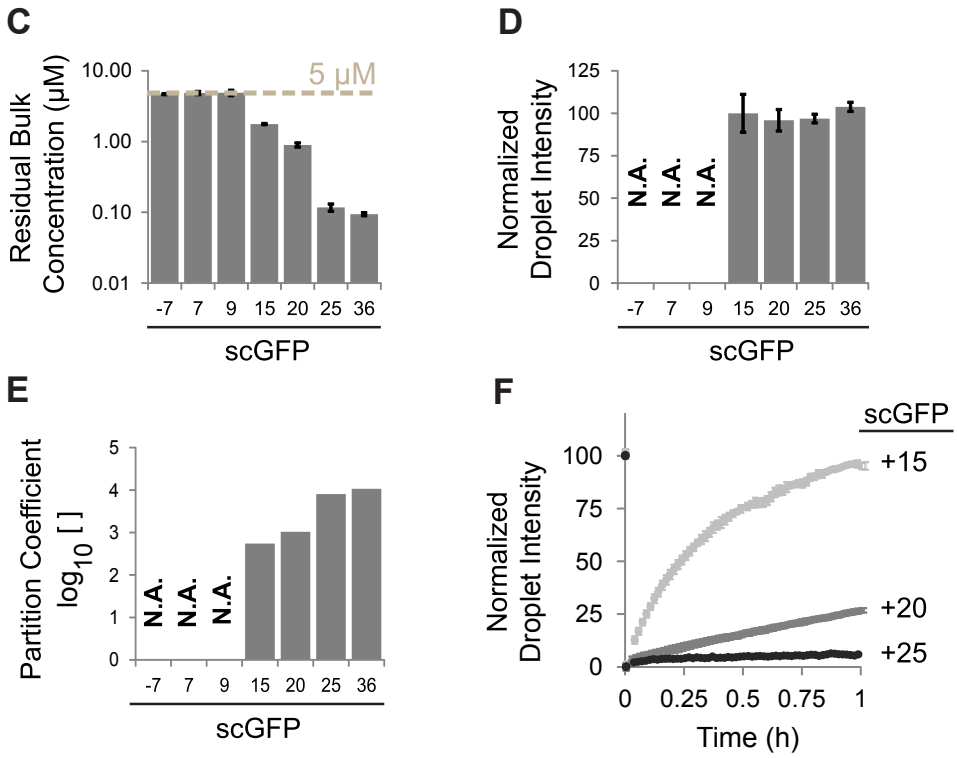

Figure 11.12: (C) Solutions containing 5 µM wild-type GFP (-7) or scGFPs (+7, +9, +15, +20, +25, +36) plus 5 µM NICD were clarified by centrifugation, and the concentration of supercharged GFP in the supernatant (residual bulk) was quantified. (D) Fluorescence intensities of phase separated droplets formed with different scGFPs. (E) Partition coefficients of scGFPs (droplet intensity / residual bulk concentration), 5 µM, in the presence of 5 µM NICD. (F) Fluorescence recovery after photobleaching analysis of scGFPs in phase separated droplets formed with NICD. In panels C, D, and F, data are represented as mean ± SEM.

11.3.4 Charge Patterning of NICD Affects Nuclear Body Formation

We next examined how phase separation is affected by the linear patterning of charged residues in NICD. In the wild type protein, many negatively charged residues are grouped into a series of clusters across the sequence (fig. 11.13). We highlight these clusters (fig. 11.13B), which we refer to as charge interacting elements (CIE), by identifying regions of at least four consecutive residues with a net charge per residue < -0.35, averaging over a sliding window of five residues.

The parameters used to define charge interaction elements were based on previous work in polymer and polyampholyte physics [126]. A five residue window corresponds to the number of residues beyond which the combined balance of chain-chain and chain-solvent interaction energy is on the order of thermal fluctuations (kT) [146, 440]. This length-scale is also referred to as a blob. A charge threshold of -0.35 corresponds to the net-charge per residue limit at which a polymer enters the strong (negative) polyelectrolyte regime on the diagram of states [126]. Finally, given a sliding window of 5 and a sliding step-size of 1, a length cutoff of four residues or longer corresponds to the length-scale at which a region will have a net charge between -0.2 and -0.4 and remain equal to or greater than the blob length scale, given the appearance of flanking residues around the element, which by definition must have a net charge of -0.2. A length cutoff of three would be too lenient, while a length cutoff of five would allow for negatively charged blob-sized regions to remain unidentified as CIEs. A cutoff of 4 residues offers an ideal compromise.

Although defining these elements represents a simplified approach for capturing the underlying electrostatic interactions, it allows us to quantify negatively charged clusters in a

consistent manner, compare among NICD charge mutants, and identify NICD-like sequences based on charge features. There are four CIEs, with a combined length of 24 residues, in wild type NICD (fig. 11.14C and D). Initial cellular studies found that NICD deletion of the amino-terminal 62 residues of NICD did not affect phase separation (fig. 11.2C). Thus we focused on charge patterning in the remainder of the protein. By shuffling charged and polar uncharged residues with respect to one another, we designed NICD charge patterning mutants that increased (charge clustered mutant or CC) or decreased (charge scattered mutant or CS) the local charge density, which increase the CIE count and the total number of residues contained within CIEs (fig. 11.13 and 11.14), while maintaining overall amino acid composition and hydrophobic/hydrophilic patterning.

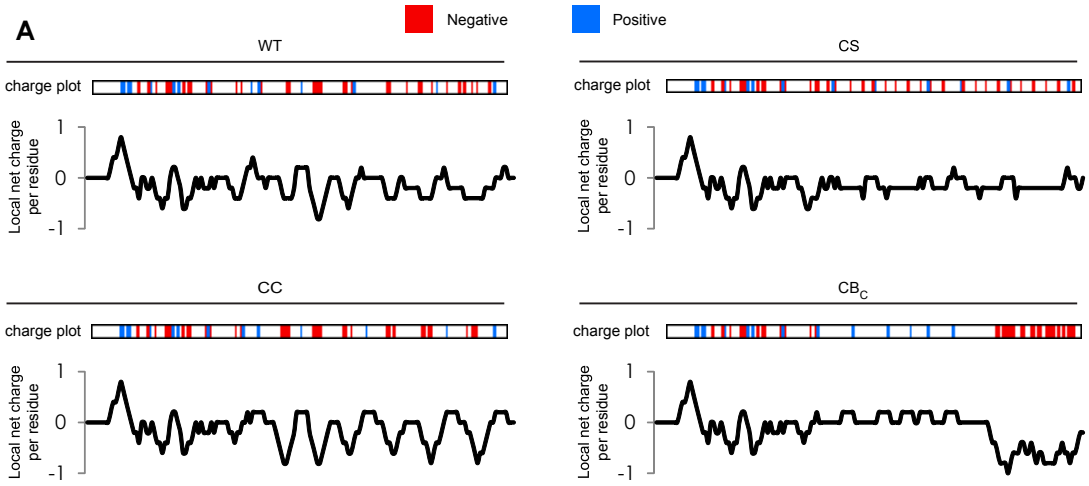

Figure 11.13: (A) The positions of negatively (D, E; red) and positively (K, R; blue) charged residues (charge plots) and linear charge density scores are shown for wild-type (WT) NICD and charge mutants (CC, CS, CB_C).

Mutant NICD proteins all expressed at similar levels in HeLa cells, and our qualitative conclusions are not dependent on expression level. Wild type NICD formed nuclear bodies in ∼70% of transfected cells (a value we normalized to 100% for comparison of different

sequences). The CC mutant formed nuclear bodies in more cells than WT NICD (fig. 11.14E), whereas the CS mutant formed nuclear bodies in significantly fewer (Figure 11.14E). Similarly, *in vitro* measurements of the saturation concentration, with scGFPs between +15 and +36, found the CC and CS mutants phase separated at lower and higher concentrations than WT, respectively (fig. 11.14F). Thus, in cells and *in vitro*, the phase separation of NICD correlates with sequence regions possessing high local charge density.

We also generated shuffled sequence mutants that clustered negatively charged residues into one region of NICD. These mutants are distinguished by the positions of the single, large charge block (CB_N, CB_{I1}, CB_{I2}, CB_C). Three of the four CB mutants (CB_{I1}, CB_{I2}, CB_C) efficiently formed nuclear bodies (figs. 11.14, 11.15 and 11.16)). We examined CB_C *in vitro*, and found it to phase separate with equivalent or greater efficiency than WT NICD for a range of scGFPs (fig. 11.14F, G; and fig 11.16C). Thus, multiple clusters of negative charges distributed across the NICD sequence are not strictly necessary for complex coacervation *in vitro* or in cells. However, it appears that the total negative charge accumulated within CIEs reflects the determinants of complex coacervation and formation of nuclear bodies by NICD mutants. Specifically, more or denser clusters of negative charges appear to lead to stronger interactions with macro-cations, enabling the charge neutralization and non-covalent crosslinking needed for complex coacervation.

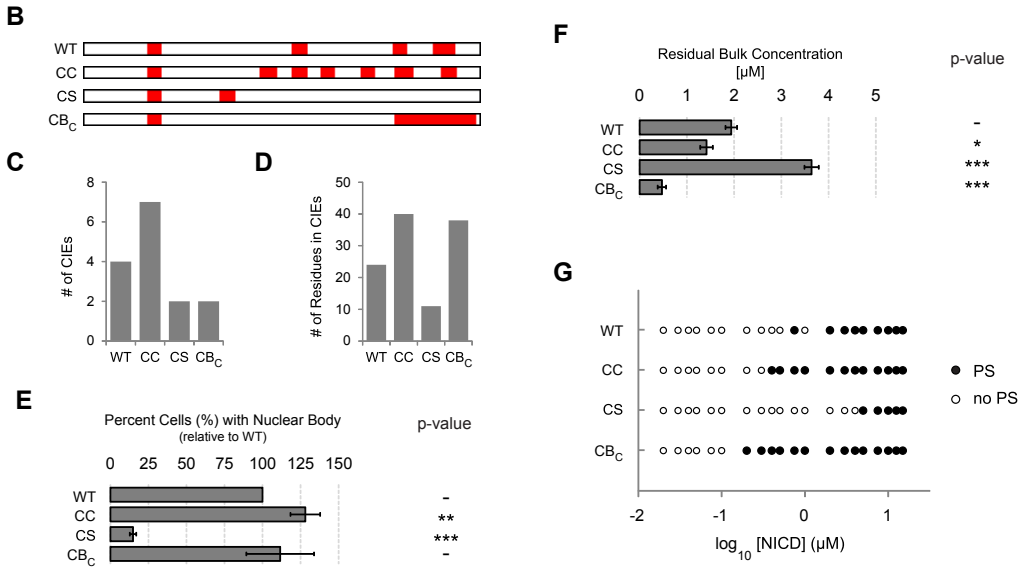

Figure 11.14: (B) The positions and length of charge interaction elements (CIEs; red) are shown for WT NICD and charge mutants (CC, CS, CB_C). (C) The number of CIEs in each sequence. (D) The number of residues in CIE regions for each sequence. (E) Normalized (to WT) percent of HeLa cells containing nuclear puncta when expressing wild type or mutant NICD. (F) The concentration of scGFP remaining in the supernatant (residual bulk) after clarification of solutions containing 5 µM NICD proteins plus 5 µM scGFP(+15). In E and F, data are represented as mean ± SEM and p-values for comparison to WT NICD represent: $* < 0.05, ** < 0.01, *** < 0.001$. (G) A range of equimolar concentrations of NICD and scGFP(+15) were mixed and scored for phase separation (black circle: phase separated; white circle: not phase separated)

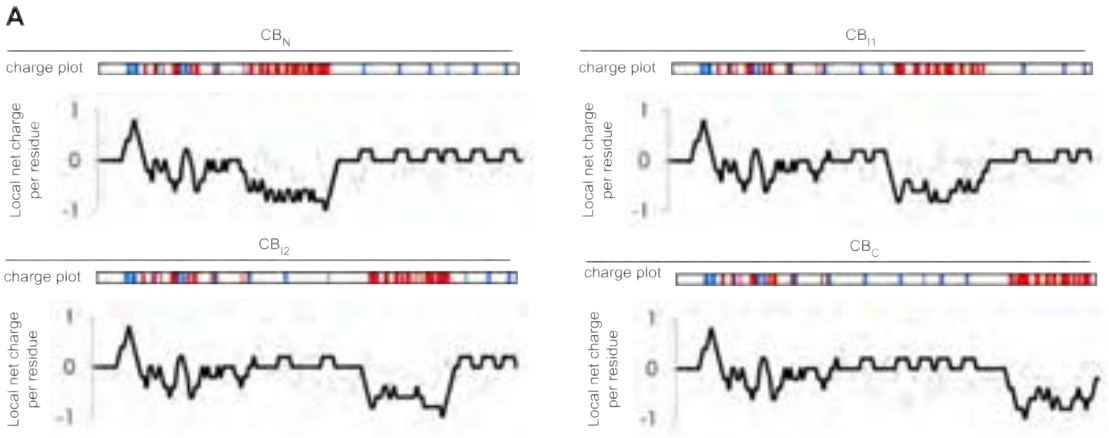

Figure 11.15: (A) The positions of negatively (D, E; red) and positively (K, R; blue) charged residues (charge plots) and linear charge density scores are shown for wild type (WT) NICD (gray dashed) and charge mutants (CB_N, CB_{I1}, CB_{I2}, CB_C; black solid). All proteins have the same overall amino acid composition, and thus the same net charge. In these designs the negatively charged amino acids in residues 61-165 were combined into a single block located at different positions in the protein. The phase behaviour of these designs is shown in figure 11.16.

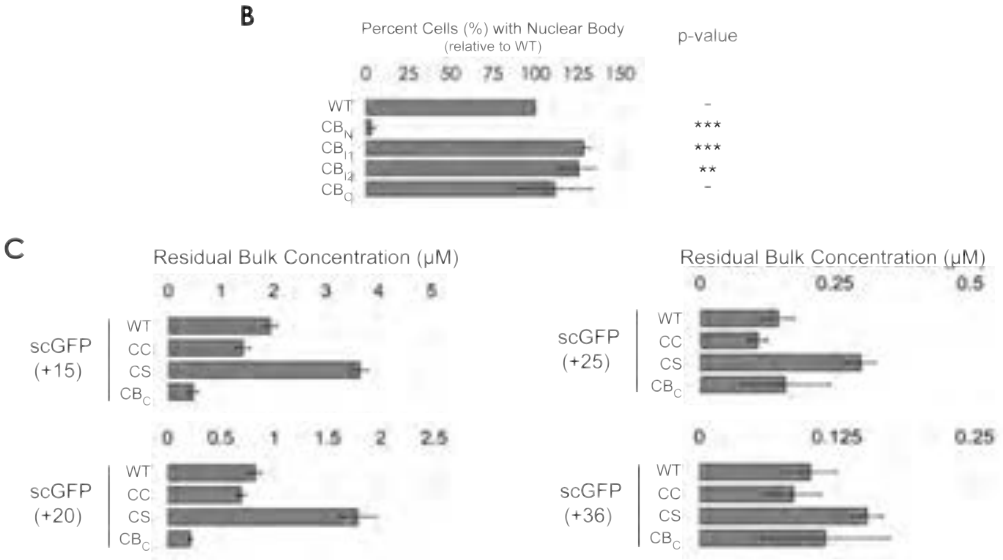

Figure 11.16: (B) Quantification of nuclear body formation for NICD charge mutants. Three of the four CB mutants produced an equal or greater percentage of HeLa cells with nuclear bodies than did WT NICD, indicating a single charge block is sufficient. Data are represented as mean ± SEM and p-values for comparison to WT NICD are: $* < 0.05$, $** < 0.01$, $*** < 0.001$. (C) NICD charge mutants (CC, CS, CB_C; 5 μM), with the same net charge but with different charge patterning, were mixed with supercharged GFPs of increasing positive charge (5 μM) and the residual bulk concentrations of scGFP were quantified. Increasing net positive charge of scGFP promotes phase separation. Residual bulk concentration for WT NICD decreases ∼20-fold over the range of positive charges tested (+15 to +36). With the same supercharged GFP species, residual bulk concentrations for NICD mutants with higher local charge density (CC and CB) were always lower than or equal to those for WT NICD. The NICD mutant with lower local charge density (CS) behaved oppositely, with higher bulk concentration than WT NICD. Data are represented as mean ± SEM.

11.3.5 Specific Residue Types Promote Formation of Nuclear Bodies in a Sequence-Independent Fashion

To identify other determinants of phase separation in NICD, we deleted a series of 6-12 amino acid segments (fig. 11.18A). Nearly all of these deletions reduced the formation of nuclear bodies (fig. 11.18B). We also generated mutants wherein different pairs of single deletions were combined (fig. 11.17). With only a single exception ($\Delta 1/\Delta 2$), double deletions led to a greater reduction in nuclear body formation than the single deletions. These data suggest that multiple elements distributed throughout the NICD sequence contribute to phase separation.

Figure 11.17: Normalized (to WT) percent of HeLa cells containing nuclear puncta (left) when expressing constructs deleted for multiple sequence elements (schematically illustrated at right).

We next examined whether these elements contribute through specific sequences of amino acids, as is typical of most protein-protein interactions, or through more general physical properties that arise from amino acid content, as is more typical of polymer-polymer interactions. We shuffled the residues in individual functionally important regions, while maintaining overall amino acid composition. For each region, we generated multiple shuffled variants (fig. 11.19), and compared their ability to form nuclear bodies to that of the

native protein and the corresponding deletion mutant. We found that every shuffled variant formed nuclear bodies nearly as well as (and occasionally better than) the native NICD (fig 11.18C). Thus, regions of NICD that promote intracellular phase separation act in a sequence-independent fashion.

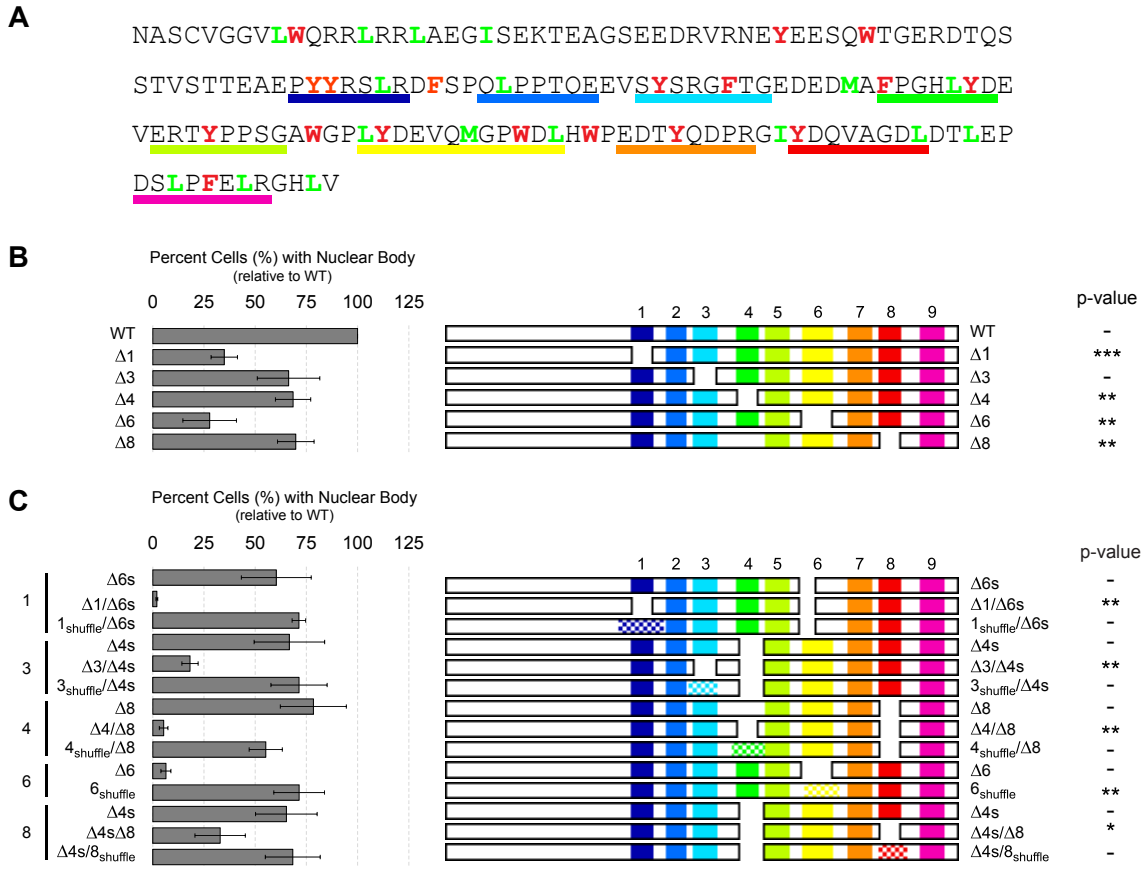

Figure 11.18: Identification of residue types that promote nuclear body formation. (A) Amino acid sequence of NICD. Aromatic and hydrophobic residues are colored red and green, respectively. Regions deleted in panels B, C, and E are underlined in matching colors. (B) Normalized (to WT) percent of HeLa cells containing nuclear puncta (left) when expressing constructs deleted for individual sequence elements (schematically illustrated at right). (C) Quantification of nuclear body formation for locally shuffled sequences. The positions of shuffled sequences are checkered.

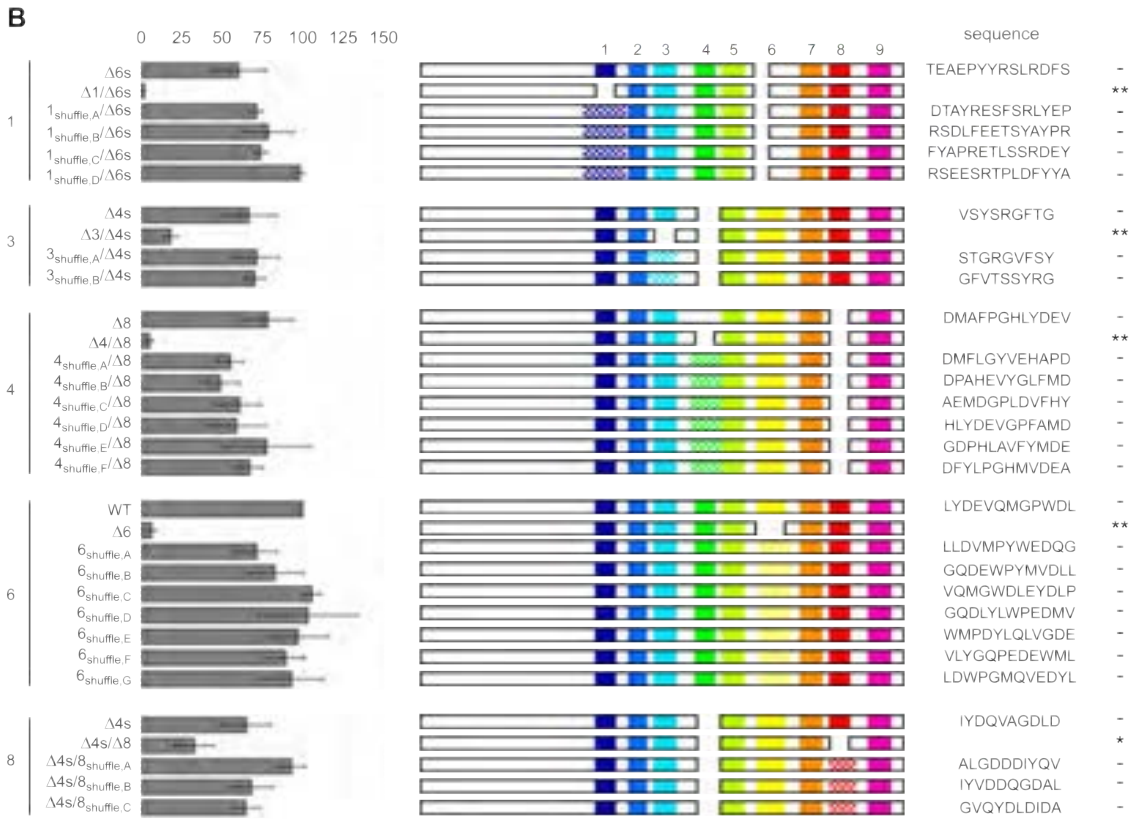

Figure 11.19: Quantification of nuclear body formation for additional locally shuffled sequences. The positions of shuffled sequences are indicated (checkered region), and the native sequence (top of each group) and shuffled sequence are shown. Local shuffle had a minimal effect on the ability of variants to phase separate

Previous studies showed that mutation of Tyr and Phe residues in IDRs from Ddx4 and BugZ inhibit their phase separation [266, 421]. Relatedly, mutation of Tyr residues in Fus and an IDR from hnRNPA2 prevented partitioning into RNA granules and phase separated droplets, respectively [282, 659]. These studies focused specifically on aromatic residues because of their striking enrichment along with Ser, Gly and Gln residues in the IDRs of various RNA binding proteins [217, 282, 421]. NICD is not enriched in any one amino acid

type, and has an amino acid distribution more typical of "generic" IDRs (fig. 11.23). We thus asked whether any amino acid type contributed more significantly to phase separation, making no a priori assumptions regarding which residue may be of interest. We addressed this issue through an unbiased statistical analysis of the residue types that are lost in each of the deletions and the effects of these losses on nuclear body formation. We first assessed all possible combination sets of residue types (NICD contains 16 unique residue types) for the correlation between the number of residues from a set lost upon each deletion and the fraction of cells with puncta associated with that deletion. For the sets of residues that correlated most strongly (critical residue sets), we then determined the enrichment of each residue type within each set relative to a random prior in a set-size matched manner (fig. 11.20). Heatmap values greater than one indicate relative enrichment. Deletion of Tyr correlated most strongly with the loss of NICD bodies, followed by Arg, Leu, Met and Trp, and Asp (figfig. 11.20). Similarly, Phe and Arg residues have been implicated as central players in the phase separation via simple coacervation of Ddx4, with Phe residues suggested to mediate intermolecular cation-pi interactions [421]. Leu and Met residues may facilitate short-range homo- and heterotypic hydrophobic interactions, while Asp is expected to drive complex coacervation through complementary electrostatic interactions with cationic partners.

To test these correlations, we mutated aromatic or hydrophobic residues at several positions in the NICD sequence. In each case, the mutations reduced nuclear body formation (11.21E). Some point mutants had a greater effect than deletion of the corresponding residues and their neighbours (compare to deletions shown in fig. 11.20 and 11.21), likely because of the introduction of positively charged Lys residues, which also decrease negative charge density. These data support our model that aromatic and hydrophobic residues play important roles in increasing the driving force for phase separation.

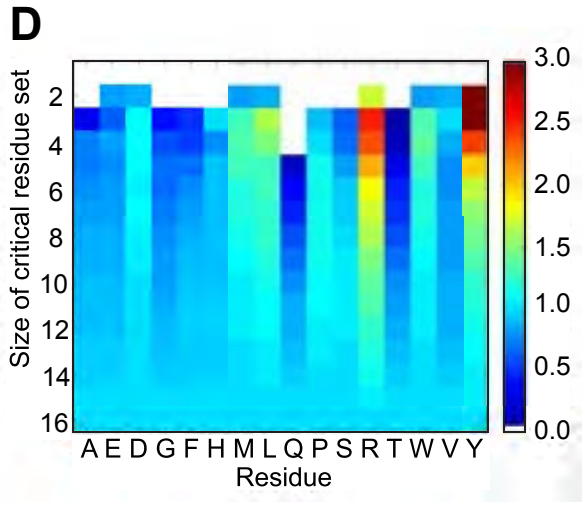

Figure 11.20: (D) Heatmap of relative enrichment of specific residue types in deletions that correlated strongly with decreased cellular puncta. Warmer and cooler colors indicate enrichment and depletion, respectively, in critical residue sets.

11.4 Discussion

NICD forms nuclear bodies when expressed in mammalian cells. Our cellular and biochemical data indicate that NICD phase separates via complex coacervation, requiring interactions with positively charged partners to neutralize its appreciable negative charge. A combination of charge neutralization, which is governed by the local linear charge density, and interactions involving aromatic and hydrophobic residues appear to be the main drivers of NICD phase separation. The local charge density of NICD and the surface charge density of its counterion partners govern the saturation concentration for phase separation and the physical properties of droplets. Statistical analysis of deletions and mutagenesis also indicate that aromatic (Tyr, Trp) and hydrophobic (Leu, Met) residues, which are distributed across the NICD sequence, contribute to phase separation. The overall amino acid composition coupled to

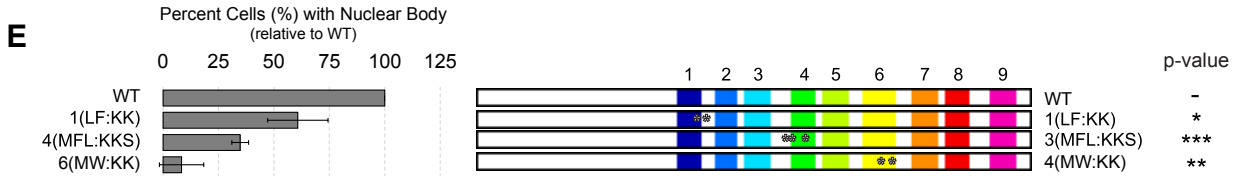

Figure 11.21: (E) Nuclear body formation by NICD point mutations. Positions of mutations are indicated by white asterisks. In panels B, C and E, data are represented as mean ± SEM and p-values for comparison to WT NICD, respective single deletion mutants and WT NICD, respectively are: $* < 0.05$, $** < 0.01$, $*** < 0.001$).

charge patterning seems to suffice for driving phase separation, i.e., the precise sequence of NICD does not appear to matter.

Our observations lead to a mechanistic model for NICD phase separation that is based on a hierarchy of interaction ranges and strengths (fig. 11.24). Long-range electrostatic repulsions among negatively charged interaction elements must be weakened in order to draw NICD into dense droplets. The dense phase is further stabilized by short-range interactions involving aromatic and hydrophobic residues. Electrostatic interactions with positively charged partners appear to neutralize the negative charge of NICD molecules and serve as non-covalent crosslinks (fig. 11.6D and 11.22). *In vitro* we have observed NICD phase separation promoted by scGFP proteins and oligoArg peptides. In cells, it is likely that NICD has many partners, with varying degrees of charge and stoichiometry of interaction, that collectively promote phase separation in the nucleus. Notably, many RNA- and DNA-binding proteins are localized to the nucleus and are highly basic - we speculate that these could play a role in mediating complex coacervation. The strength of complementary electrostatic interactions and their contribution to phase separation increase with total positive charge on scGFP/oligoArg, and are modulated by the charge distribution along the NICD. Clustering

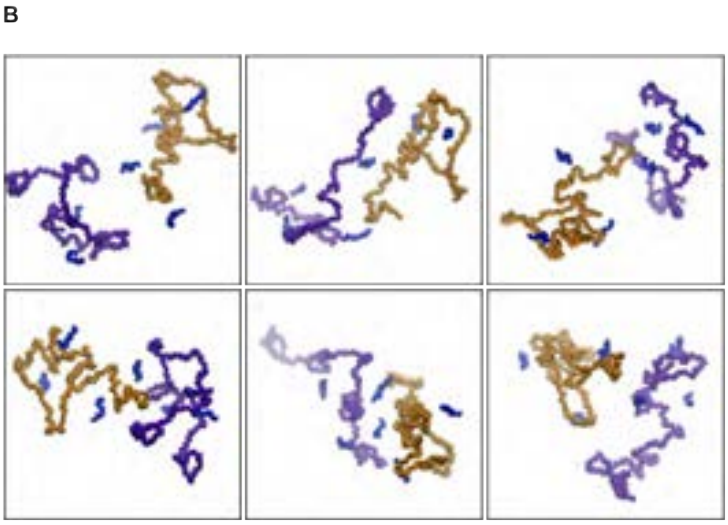

Figure 11.22: Representative snapshots of pairs of NICD molecules from atomistic Monte Carlo simulations in the presence of Arg_6 peptides. Atomistic Monte Carlo simulations were performed using the ABSINTH implicit solvation model and forcefield paradigm. Backbone-only representations are shown, with the two NICDs shown in magenta and orange and Arg_6 peptides shown in blue.

Asp/Glu residues to produce one or more blocks with high local charge density promotes phase separation, whereas distributing these residues more evenly attenuates phase separation (see fig. 11.13 and 11.15). The exception here is the CB_N variant, which attenuates the driving force for phase separation. Although this warrants further investigation, atomistic simulations suggest that CB_N is more collapsed than the other CB mutants. This collapse originates from intramolecular electrostatic contacts between the negative charge block and a region in the N-terminus. This might limit the accessibility of CIEs to multivalent counterions thus weakening the driving force for phase separation via complex coacervation. In effect, the intramolecular interactions are able to effectively compete with inter-molecular interactions.

In addition to charge, Tyr, and to a lesser degree Leu, Met, and Trp residues also contribute to NICD phase separation. Interactions of hydrophobic residues are inherently short range, and mediate multivalent adhesions among NICD molecules. These contacts are weaker analogs of domain-ligand interactions that promote phase separation of modular signaling proteins [29, 30, 188, 330]. The importance of these residue types in promoting phase separation and/or partitioning into the droplet phase is also suggested by mutagenesis of aromatic residues in the IDRs of Ddx4, FUS and BugZ [217, 267, 282, 420, 421, 659]. Rather than mutating specific residues, we approached the issue agnostically, making deletions across the sequence and then using statistical analyses to identify functionally important residue types. Moreover, unlike Ddx4 and FUS, which form cellular bodies as part of their normal functions, NICD has (as far as we know) not evolved to form nuclear bodies. Nevertheless, we arrive at a similar conclusion, i.e., that Tyr is especially enriched in segments whose deletion is most deleterious to intracellular phase separation. This suggests that aromatic residues may play an important role in promoting the phase separation of disordered proteins in general, and their enrichment in the IDRs of RNA binding proteins may have a physical basis.

No single linear motif or specific amino acid sequence in NICD is crucial for driving phase separation. Instead the overall amino acid composition combined with the patterning of negatively charged residues and the distribution of aromatic residues along the linear sequence appears to be important. This indicates that adhesive structures form through non-specific interactions, perhaps through sidechain contacts alone, or through additional backbone contacts such as cross-beta strands. Thus, once drawn together by charge neutralization, NICDs appear to form labile complexes in which weak aromatic/hydrophobic and electrostatic interactions are made and broken rapidly, imparting liquid-like behaviour to the phase separated state.

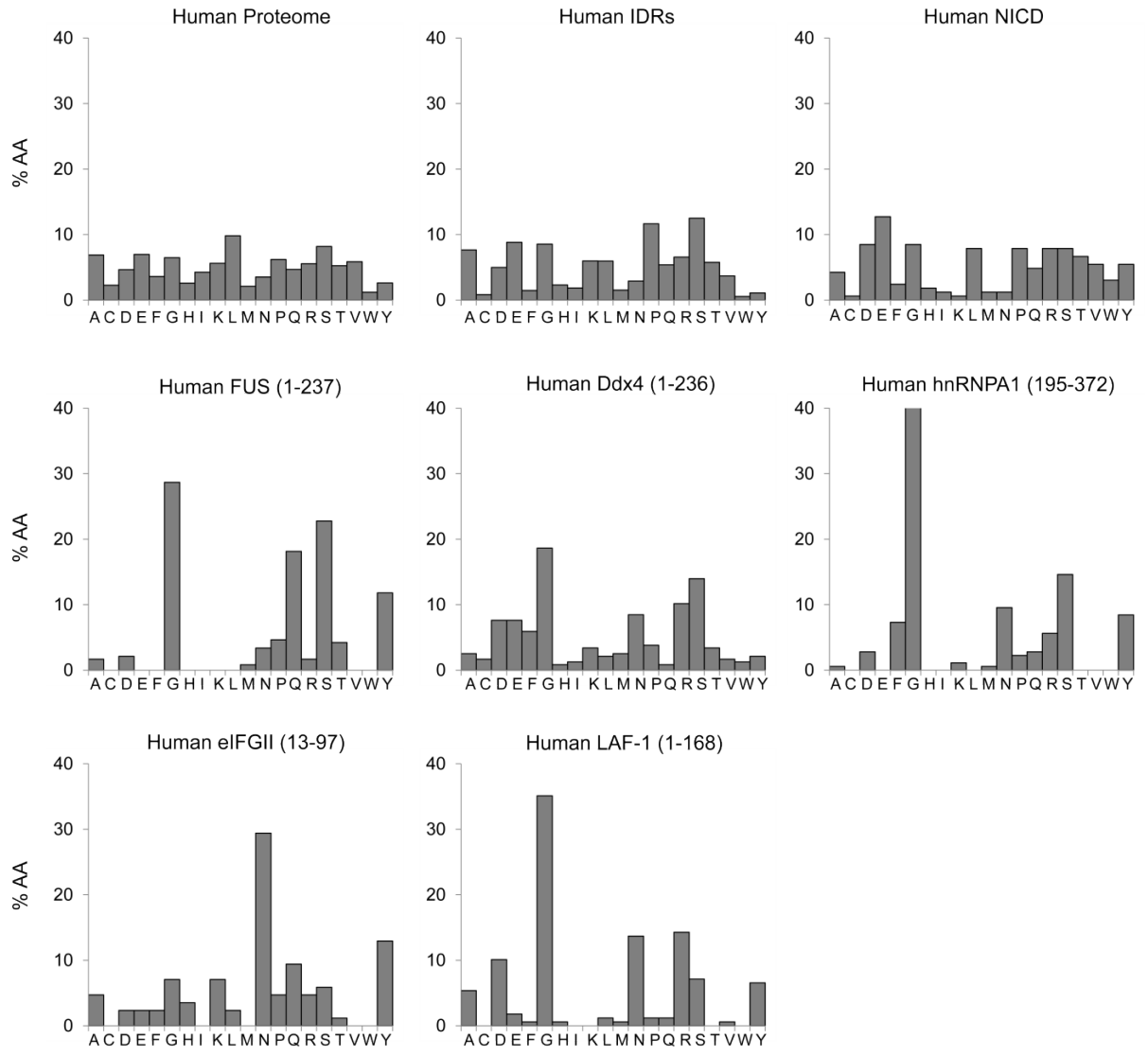

Figure 11.23: The distribution of amino acids (AAs) for various groups of sequences were analyzed, including for the entire human proteome (UniProt: UP000005640_9606), IDRs in the human proteome predicted by the MobiDB consensus prediction, and several IDRs (NICD, FUS (1-237), Ddx4 (1-236), hnRNPA1 (195-372), eIFGII (13-97), Laf-1 (1-168)) shown previously to undergo phase separation. The sequence of NICD is not low complexity and its amino acid distribution is more typical of the average human IDR

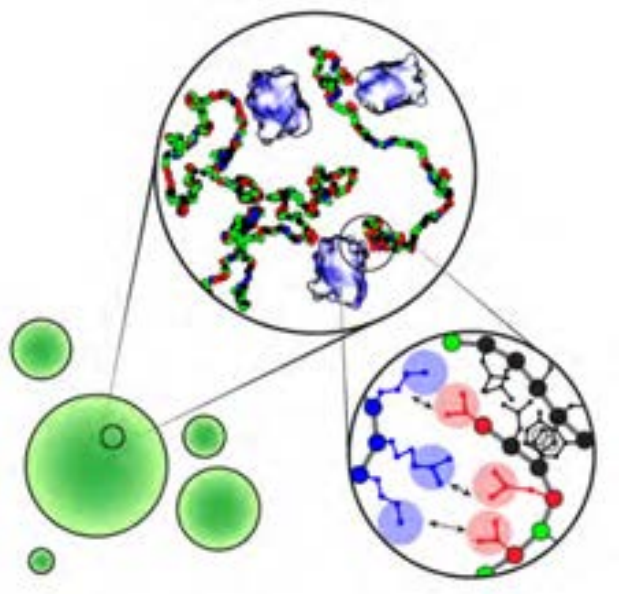

Figure 11.24: Model depicting the hierarchies of interactions that drive NICD phase separation via complex coacervation.Structures of phase separated droplets on three length scales, micrometer (bottom left), nanometer (top center), and atomic (bottom right). On the micrometer scale NICD forms liquid-like spherical droplets. Phase separation of NICD requires a multivalent counterion such as the positively charged supercharged GFP (scGFP). On the nanometer scale NICD is depicted as chains with a single bead per residue, which contact one another indirectly through scGFP counterions that likely bind the negatively charged clusters along the contour of NICD and directly through aromatic / hydrophobic interactions. In this representation, Asp / Glu residues are shown in red, Lys / Arg in blue, polar residues in green, and aromatic / hydrophobic residues in black. The scGFP(+36) molecules are shown in electrostatic surface representations. Dark blue patches indicate high positive surface charge. At the atomic scale, complementary electrostatic interactions as well as interactions involving aromatic (Tyr) and hydrophobic (Leu) residues physically crosslink NICD molecules. Note that the structural nature of these interactions remains unknown and should not be inferred from schematic image.

Recent studies of IDRs that phase separate have focused on two distinct archetypes. The first is enriched in polar residues such as Gly, Ser, Gln, and Asn, as in the IDRs of FUS and hnRNPA1. The second is enriched in charged residues, but the overall net charge per residue is low, and oppositely charged residues are segregated into blocks along the linear sequence, as in the IDRs of Ddx4 and LAF-1. Both archetypes phase separate via simple coacervation, and do not appear to require heterotypic interactions with a partner. NICD represents a third archetype. It is similar to Ddx4 and LAF-1 in that roughly a quarter of its residues are charged and the charges are clustered into blocks. However, NICD has a substantial net negative charge with more than twice as many negatively charged residues as positively charged ones. These properties lead to the requirement for positively charged multivalent ligands / counterions that drive phase separation via complex coacervation. The data to date indicate that three different strategies, based on the three archetypes of IDRs, can promote intracellular phase separation.

We analyzed the human proteome (UniProt: UP000005640_9606) using the consensus disorder prediction database MobiDB 2.0 to identify long ($>$ 100 residue) IDRs with sequence properties similar to NICD. We sought IDRs with a fraction of charged residues > 0.25 and a two-fold excess of negatively charged residues over positively charged residues. We also required that at least twenty residues be encompassed by CIEs. Based on this analysis we identified 464 unique NICD-like IDRs from 443 unique proteins. Further filtering of this list for sequences that had tyrosine and/or leucine content $\geq$ 6 % of total residues in disordered regions identified 260 unique proteins. Performing Gene Ontology (GO) molecular function over-representation (experimental only) using PANTHER (Release 2015-08-06) on these two sets identified significant enrichment in GO terms associated with nucleic acid binding and regulation of nucleic acid biosynthesis [382]. Furthermore, a number of these

proteins are annotated as being localized in nuclear body structures, particularly the nucleolus. It is possible that some of these proteins contribute to the formation of nuclear bodies through complex coacervation. In addition, since many positively charged proteins bind to RNA/DNA, it is possible that complex coacervation driven by NICD-like IDRs could compete with complex coacervation of RNA/DNA binding proteins and nucleic acids. This suggests the possibility that NICD-like IDRs might be effective at dissolving ribonucleoprotein bodies and/or sequestering molecules that otherwise partition into ribonucleoprotein bodies, as NICD appears to do with nuclear paraspeckles (fig. 11.4B).

We used NICD as a model system to understand the phase separation of an archetypal disordered protein. However, these findings also have potential implications for the biology of Nephrin. In mammals, Nephrin is a transmembrane adhesion receptor expressed primarily in podocyte cells of the kidney. Tyrosine phosphorylation of the intracellular domain of Nephrin, and consequent binding of Nck and assembly of cortical actin, are necessary for proper formation of the filtration barrier of the kidney [49, 269, 270]. Interactions between phospho-Nephrin, Nck and N-WASP result in Nephrin phase separation into membrane attached puncta *in vitro* and in cells (S. Kim and M.K. Rosen, unpublished) [29]. Assembly of these puncta has been attributed to multivalent SH2-phosphotyrosine and SH3-polyPro interactions. However, the SH2 domain, the first SH3 domain, and an adjacent linker region of Nck are basic (predicted pI ∼9). N-WASP also has a poly-basic region. It is possible that these basic elements act analogously to scGFP or oligoArg peptides to promote Nephrin phase separation. Thus, non-specific charge-mediated interactions, along with specific modular domain interactions, could contribute to the formation of membrane puncta by the ternary Nephrin/Nck/N-WASP system. The disordered cytoplasmic regions of a number of transmembrane signaling proteins, including LAT, FAT1, and PDGFR, are also highly acidic, and interact with adaptor proteins that also have basic elements [563]. Thus, it is

possible that complex coacervation could also contribute more broadly to the cytoplasmic clustering and intracellular phase separation of membrane-anchored proteins.

Of particular interest is the Interferon alpha/beta receptor 2 which contains a $\sim$100 residue region (316 - 419) that has almost identical sequence features to NICD. This region is a sub-region within a larger intracellular domain on the cytoplasmic side of a transmembrane domain (also analogous to Nephrin). This Interferon alpha-beta receptor 2 intracellular domain (IR2-ICD) contains all of the features associated with nephrin to a higher degree, many more tyrosines (which undergo phosphorylation, many large acidic patches, and is directly involved in signal transduction (STAT1/STAT2/STAT3 activation) through JAK1 activation [152, 422]. We speculate that complex coacervation could provide a mechanism for signal amplification, whereby phosphorylation of the IR-ICD provides a mechanism for signal attenuation.

Chapter 12

Phase Behaviour of Disordered Proteins Underlying Low Density and High Permeability of Liquid Organelles

The following section is taken from the paper **Phase behavior of disordered proteins underlying low density and high permeability of liquid organelles** by M-T. Wei*, S. Elbaum-Garfinkle*, A.S. Holehouse*, CC-H. Chen, M. Feric, C.B. Arnold, R.D. Priestley, R.V. Pappu, C.P. Brangwynne (* denotes co-first authors). This was published online in *Nature Chemistry*, in May 2017. All experimental work detailed in this chapter was performed by Steven Wei and Shani Elbaum-Garfinkle, while much of the theoretical analysis associated with those data and all simulation work was performed by A.S.H. Details associated with the experimental work are included for clarity and completeness. The text has been expanded to include additional detail absent from the paper.

12.1 Background

Living cells consist of thousands of different proteins, nucleic acids, lipids, and small molecules, held together by a phospholipid membrane. They integrate extrinsic and intrinsic stimuli over a wide range of length-scales and time-scales to produce behaviour than can adapt and respond to their environment in a spatiotemporal manner. For unicellular organisms, the goal of this adaptation is realized in terms of the survival of an individual, which may be mediated through apparently 'selfish' behaviour, or though collective, population level behaviour that while apparently 'altruistic' has the emergent effect of statistically improving an individual's fitness. For multicellular organisms, individual cells function as part of a tightly coupled and highly dependent network. Collectively, that network abstracts certain tasks to specific cell types, allowing substantially enhanced environmental specificity, improved collective efficiency, and greater individual robustness, all at the expense of environmental plasticity.

Given the incredible complexity faced by both multicellular and unicellular organisms, a collection of fundamental questions in biology revolve around cellular organization. How do cells organize themselves to facilitate their evolutionary goals? In a more abstract sense, this question could be cast in terms of information how do cells most efficiently convert energy into information, and how do they ensure that information remains plastic, accessible, yet durable?

To organize their contents, cells construct a range of different intracellular organelles that localize distinct sets of molecules, allowing spatiotemporal control of molecular interactions. In addition to canonical vesicle-like organelles (which we define here as organelles that are

surrounding by distinct barriers, typically a phospholipid membrane) there are dozens of non-membrane bound, RNA and protein rich organelles within the cell nucleus and the cytoplasm [66, 217, 282, 443, 551]. Despite their lack of an enclosing membrane, these organelles are able to concentrate molecular components and play important roles in key intracellular functions such as RNA transcription and processing, and in the regulation of protein translation.

It is now recognized that membraneless organelles, including P granules, nucleoli, and stress granules, are condensed liquid-like droplets of RNA and protein that form via phase separation. Indeed, many such organelles exhibit classic signatures of liquids, including rapid exchange dynamics of their contents with their surroundings, spherical shapes, coalescence upon contact, and flowing and dripping in response to shear stresses [65, 66]. These properties allow membraneless organelles to concentrate molecular reactants while maintaining fluidity to facilitate interactions among the constituent molecules. A growing number of studies have demonstrated the liquid-like nature of membraneless organelles and the relevance of liquid-liquid demixing as a fundamental physical mechanism explaining their formation [41, 67, 320, 330]. There is also increasing support for a link between the material properties of membraneless organelles and cell physiology as well as disease states [162, 338, 399, 406, 443, 668].

A number of key questions remain unanswered regarding the physicochemical driving forces for phase separation and the macromolecular organization within membraneless organelles. Intrinsically disordered proteins or regions (IDPs/IDRs) are, in many cases, the drivers of phase-separation that give rise to membraneless organelles, although how protein disorder contributes to phase separation remains unclear [7, 162, 399, 420, 421]. In *C. elegans*, germ line P granules are RNA and protein rich droplets that are implicated in specification of germ cells. P granule assembly is driven by several proteins with IDRs [618, 625]. These proteins

including LAF-1, an abundant DDX3 family protein which contains an arginine/glycine-rich (R/G or RGG) domain that is necessary and sufficient for phase separation [162]. In addition to LAF-1 PGL-3 is another P granule protein which contains RGG domains that are important for phase separation [510]. R/G-rich IDRs are also found in the nucleolar protein FIB-1, which drives assembly of a core droplet within the nucleolus [41, 172]. Other examples include WHI3, which contains a Q-rich IDR and drives the formation of liquid-like puncta in the cytoplasm of fungi [668]. Similarly, the stress granule proteins hnRNPA1 and FUS contain IDRs and are also involved in neurodegenerative diseases [338, 399, 443]. The molecular concentration within such droplets is expected to influence behaviors including molecular sequestration, promotion of various reactions, and the nucleation of amyloid-like fibers that are associated with disease [16, 338, 399, 436, 443].

Despite the importance of intra-droplet concentration, there are difficulties associated with measuring the full coexistence curves (i.e., binodals) that define the protein concentrations inside and outside of the droplet. As a result, little is known about how droplet properties emerge from the underlying RNA/protein interactions Aumiller2016-fw. The prevailing wisdom surrounding the intra-droplet protein concentration is that these organelles are a similar density to that consistent of a polymer melt, with protein concentrations in the 100-500 mg/ml range. These numbers are in part motivated by qualitative estimates of the density, and in part through simple polymer theories fit to the low concentration arms of the binodal curves using Flory-Huggins based theoretical models.

In this work, we utilize a novel method based on fluorescence correlation spectroscopy (FCS) measurements to infer second virial coefficients, molecular diffusion coefficients, and binodals for LAF-1 and its intrinsically disordered RGG domain in the presence and absence of RNA molecules. By combining these measurements with a theoretical framework and insights from

atomistic simulations, we uncover a rich physical picture of the interactions that underlie the phase behavior and properties of LAF-1 droplets. These results show that intra-droplet concentrations are surprisingly low, and suggest that condensed phases are akin to semidilute polymer solutions. Large-scale conformational fluctuations originating from the intrinsically disordered RGG domain in LAF-1 are critical for the formation of such low-density droplets. We also determine a structural length scale, the mesh size, which characterizes the molecular organization within droplets both *in vitro* and *in vivo*. The inferred mesh sizes of P granules in living *C. elegans* embryos, as well as other membraneless organelles, suggest the broader relevance of our findings for droplets within living cells.

12.2 Methods and Results

12.2.1 Phase Separation in LAF-1 is Driven by the RGG Domain

It had previously been demonstrated that the RGG domain is necessary and sufficient to drive phase separation [162]. We analyzed the amino acid sequence of LAF-1 to build a molecular picture of the various local regions within the protein. This analysis was performed using the localCIDER software package (see chapter 4)and is shown in fig. 12.1.

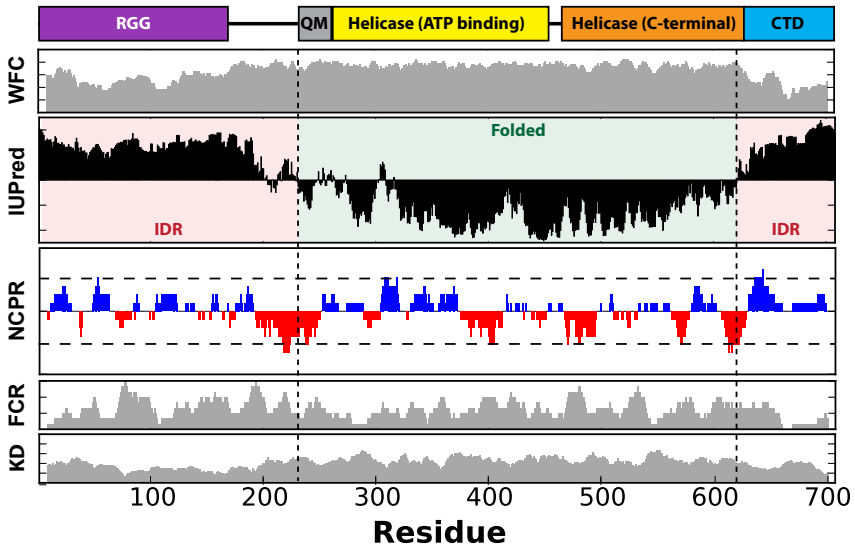

Figure 12.1: Linear sequence analysis of LAF-1. Each track describes a difference type of sequence feature to provide a general summary of the linear amino acid sequence

The analysis reveals two well defined and a two-domain helicase. WFC represents Wootton-Federhen sequence complexity; the RGG and CTD domains show a significantly reduced complexity when compared to the remainder of the sequence. IUPred represents the predicted disorder score based on the IUPred algorithm; N-terminal and C-terminal intrinsically disordered regions (IDRs) are identified. NCPR represents the linear net charge per residue;

the N-terminal domain shows both net positively charged local regions and net negatively charged regions, suggesting electrostatic interactions may play a role in driving RGG-RGG interaction. FCR represents the fraction of charged residues; several regions in the RGG contain a high FCR despite minimal net charge, indicating these regions have the characteristics of a strong polyampholyte. KD represents the Kyte-Doolite hydrophobicity scale; the folded domains are significantly more hydrophobic than the IDRs.

The RGG domain has a strongly biased sequence composition, with a high fraction of glycine, arginine, tyrosine. These amino acid properties are similar to are sequence that are known to phases separate, but appear to show a more extreme variant, with a higher fraction of glycine than many other IDRs that drive phase separation. We sought to characterize the phase behaviour of full-length LAF-1 and the RGG in isolation

Figure 12.2: LAF-1 amino acid sequence. The RGG domain is bolded and underlined (and represents residues 1-168)

12.2.2 Ultrafast-Scanning FCS Measurements of Coexistence Curves

FCS is a powerful technique that relies on measuring the fluorescence intensity fluctuations of labelled molecules within small, femtoliter excitation volumes [351]. FCS allows for precise measurements of molecular concentrations and diffusion coefficients, and has been employed for studying protein aggregation and assembly [236, 456]. However, standard FCS methods have well-known limitations, including the need to calibrate the fluorescence excitation volume [487]. This calibration is achievable in a uniform aqueous medium, but can become problematic in a droplet-forming system due to refractive index variations. To overcome these limitations, we developed a novel approach, called ultrafast-scanning FCS (usFCS). This approach uses a tunable acoustic gradient index of refraction (TAG) lens placed in the back focal plane of an oil immersion objective, as shown in fig. 12.3a [155–157, 380].

The TAG lens allows axial scanning of the sample at very high frequency (70 kHz). The axial scan range (Z) is adjustable by changing the applied voltage. Importantly, because scanning is performed through the standing acoustic wave there are no mechanical parts, greatly improving the fidelity and robustness of the approach. The tunable scanning distance serves as an external ruler for measuring the unknown detection volume within droplets. This allows us to estimate the size of the measurement volume and thus determine molecular diffusion coefficients, D, from the characteristic decay times, τ_D, of measured autocorrelation functions as shown in fig. 12.4.

The samples are excited using a diode-pumped solid-state laser with emission wavelengths of 491 nm. After passing through the TAG lens, the light is focused into the sample using an oil immersion objective. The fluorescence emission is collected through the same objective, separated from the excitation light by a dichroic mirror and focused into a confocal pinhole

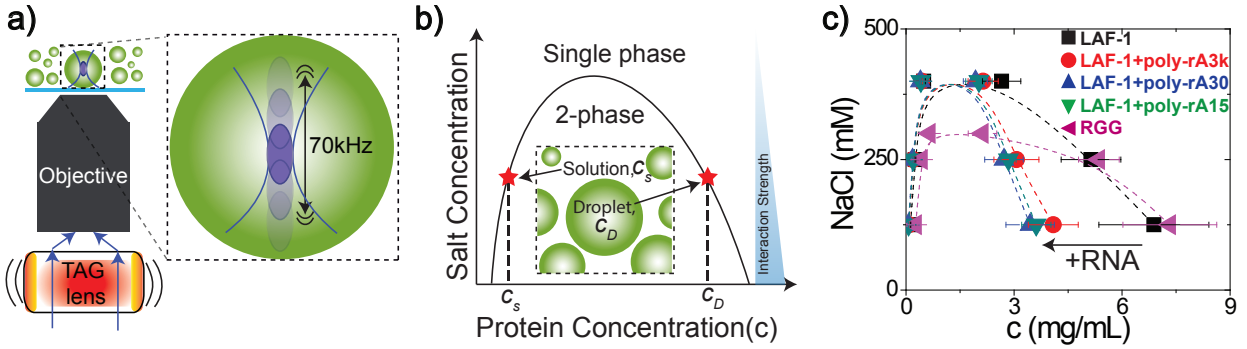

Figure 12.3: Measured binodals for the RGG domain and LAF-1. The latter is measured in the absence or presence of RNA using ultrafast-scanning FCS approach. (a) A schematic illustration of the microscope with an acoustically modulated beam that is controlled by a tunable acoustic gradient index of refraction (TAG) lens. The system focus can be axially scanned along the optical-axis at a frequency of 70 kHz. (b) Schematic showing a typical binodal with increasing protein concentration along the abscissa and increasing NaCl concentration along the ordinate. Our measurements show that the salt concentration decreases the strengths of two-body interactions for RGG domain/LAF-1 systems. (c) The measured binodals of the RGG domain as well as LAF-1 in the presence and absence of RNA.

unit. The fluorescence light is filtered by a long pass filter, photons are detected by a photomultiplier tube, and their arrival times are registered by a data acquisition card. For this study, the scanning distance was kept at $\sim 2\,\mu$m, an order of magnitude smaller than the size of the liquid droplets.

When the measurement volume is axially scanned, Z (axial scan distance), at a constant frequency f, the autocorrelation function for simple diffusion is:

$$G(\tau) = G(0)\left(1+\left(\frac{\tau}{\tau_D}\right)\right)^{-1}\left(1+\left(\frac{\tau}{\kappa^2\tau_D}\right)\right)^{-0.5}\exp\left(\frac{-(Z\sin(\pi f\tau))^2}{(2\omega)^2}\frac{1}{1+\left(\frac{\tau}{\kappa^2\tau_D}\right)}\right) \tag{12.1}$$

Here, $G(0)$ is magnitude at short time scales, τ is the lag time, τ_D is the half decay time, κ is the ratio of the axial to radial measurement volume, and ω_z is the depth of focus. The parameters τ_D, $G(0)$, and ω_z were allowed to vary and were optimized to fit the measured auto-correlation trace, while the parameters Z and f were kept fixed. ω_z provides a measurement of the confocal volume, which in turn can be used to determine the diffusion.

We used usFCS to measure concentrations within the dilute and dense phases. These concentrations correspond to the low and high concentrations arms of the binodal curves, as illustrated in figure 12.3 and described in detail in chapter 13. For a given NaCl concentration, the equilibrium protein concentration outside the droplet, c_s, defines a point on the left arm of the binodal (solid curve in fig. 12.3) whereas the protein concentration inside the droplet, c_D, defines the corresponding point on the right arm of the binodal. At 125 mM NaCl, the value of c_S for LAF-1 is 0.124 ± 0.009 mg/mL (1.5 ± 0.11 µM); the droplets that condense from solution are at a concentration of c_D = 6.88 ± 1.52 mg/mL (86.5 ± 19.2 µM), see fig. 12.3c). These values are surprisingly low, given that the folded proteins lysozyme and γ-crystallin, although fundamentally different from IDPs, are also known to phase separate, but at concentrations that are approximately two orders of magnitude higher (∼100-500 mg/mL) than what we measure here [69, 575].

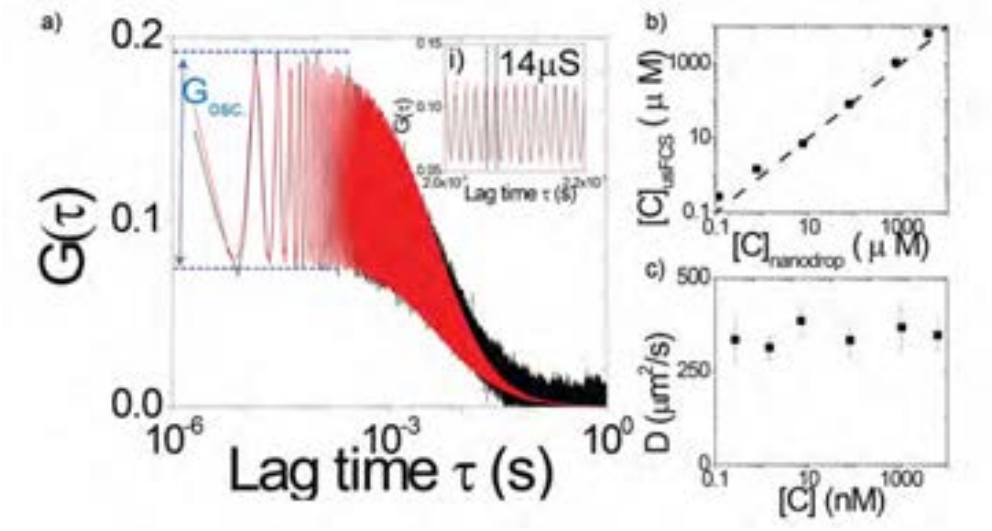

Figure 12.4: The usFCS provides a calibration free method for determining the intradroplet concentration (a) Fluorescence autocorrelation of 14 nm hydrodynamics radius polystyrene particles while scanning at frequency 70 kHz. The fit to Equation 12.1 (red line) is shown. The magnitude of autocorrelation function oscillations with 14 µs period at short time scale, G_{OSC}, depends on the ratio between axially scanned distance (Z) and depth of focus (ω_z). The inset shows fluorescence autocorrelation as a function of delay time (τ) between 2 ms and 2.2 ms. The period (T) of autocorrelation curve is 14 µs, which indicates the TAG lens scanning frequency (T^{-1} ∼70 kHz). When compared with standard FCS, ultra-fast-scanning FCS (usFCS) has several strengths. It increases the statistical accuracy for slowly moving molecules by effectively sampling a larger volume. Improving this statistical accuracy allows for shorter measurement times than standard FCS, which helps to ensure accurate correlation curves. This approach also facilitates low excitation intensity, reducing the effect of photo-bleaching and optical saturation. Here, we show both (b) molecular concentration and (c) diffusivity from usFCS measurements as a function of Dylight 488 concentration.

To ensure this was not an error associated with the usFCS setup we confirmed these low concentration usFCS measurements using different fluorescent labels (fig. 12.5a). In addition, we used an orthogonal three-dimensional confocal microscopy approach (fig. 12.5b). To calculate the droplet concentration using this orthogonal method we first used 3D confocal microscopy to measure the volume fraction of droplet in a 2-phase mixture (ϕ_D). We then took advantage of the fact that for a two-phase system the bulk (total) concentration of protein must equal the volume-fraction weighted concentrations in the dense and dilute phases, i.e.

$$C_B = C_D\phi_D + C_S(1 - \phi_D) \tag{12.2}$$

In equation 12.2, C_B is the bulk concentration, C_D is the concentration inside the droplet and C_S is the saturation concentration (i.e. the protein concentration outside the droplet). This equation can be rearranged to obtain equation 12.3

$$C_D = C_S + \frac{C_B - C_S}{\phi_D} \tag{12.3}$$

Here, C_S is the critical concentration and C_B is the total protein concentration determined using 280 UV absorption. The total protein concentration can be solved under conditions where phase separation is inhibited (such as at high NaCl) to ensure we obtain an accurate measurement. By combing the values obtained through concentration measurements and 3D confocal microscopy we can solve equation 12.3 for C_D. Using this approach, we calculated the protein droplet concentration for a variety of NaCl concentrations and compared them to

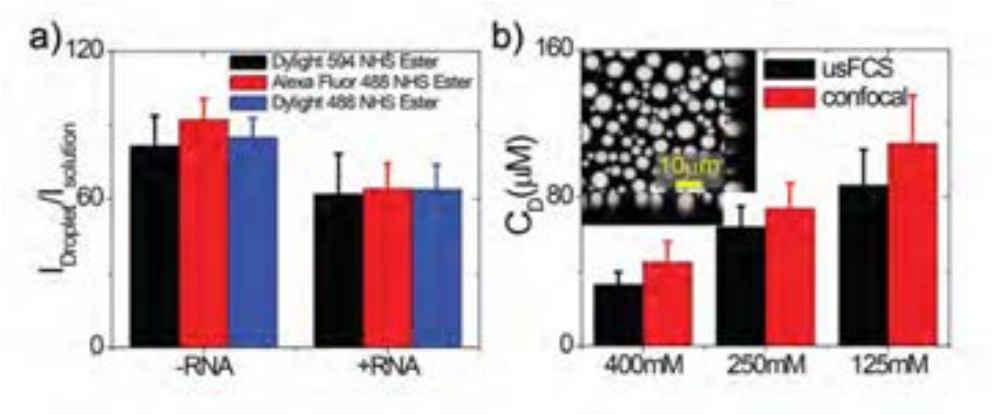

Figure 12.5: To assess the quantitative accuracy of our usFCS results, we used (a) three different fluorescent dyes to label LAF-1 and (b) determined the protein concentration using three-dimensional confocal microscopy.

values obtained via the usFCS. The results are reported in fig. 12.5, which show quantitative agreement between both methods.

Taken together, these findings demonstrate that while LAF-1 droplets are roughly 50 times more concentrated than the dilute phase, they are still at a very low concentration, which corresponds to a number density of 5×10^{-5} molecules/nm^3.

The width of the two-phase regime, quantified in terms of the ratio of c_D to c_S, decreases with increasing NaCl concentrations. This yields a concave down paraboloid binodal for LAF-1 (fig. 12.3c), which is characteristic of many polymeric systems [504]. The RGG domain of LAF-1 is necessary and sufficient to drive phase separation. Interestingly, the right and left binodal arms of the RGG domain alone are at mass concentrations that are comparable to that of full length LAF-1, although the critical NaCl concentration is lower than for full-length LAF-1. We also measured binodals in the presence of different types of generic RNA molecules, which may be expected to impact the phase diagram, since they are known to

modulate the fluidity of LAF-1 droplets. In the presence of polyadenylate RNA (poly-rA) of various lengths, the low concentration arm of the LAF-1 binodal and the concentrations corresponding to the critical region remain essentially invariant. However, upon addition of RNA we observe a marked shift of the high concentration arm of the LAF-1 binodal, toward lower values of c_D (fig. 12.3c).

12.2.3 Quantifying B_2 by usFCS

The remarkably low concentration of LAF-1 droplets must arise from the underlying protein-protein interactions, which can be modulated by RNA. Indeed, in mean field models, such as the Flory-Huggins theory, the sign and magnitude of the effective two-body interactions determine the phase behavior of polymer solutions [179, 246]. These interactions are quantified using the dimensionless Flory interaction parameter χ, and are measured in terms of molecular dissociation constants K_D or second virial coefficients B_2. We used usFCS to estimate the apparent values of B_2 as a function of NaCl concentration.

For concentrations that are below c_S, the diffusivity of a protein molecule can be influenced by interactions with other proteins [220, 581]. Interactions that are, on average, attractive will diminish the protein diffusivities, whereas two-body interactions that are, on average, repulsive will lead to larger effective diffusion coefficients (fig. 12.6). We measured the LAF-1 diffusivity as a function of LAF-1 concentration at a range of different NaCl concentrations. For all NaCl concentrations, the diffusivity measurements are made under conditions where the protein concentrations are always below the low concentration arm of the measured binodals (by up to an order of magnitude) to ensure we are in the one-phase regime. To determine B_2 from these data we use a formalism proposed by Harding and Johnson. Plotting

protein concentration (c) vs. diffusion constant (D) according to the following equation gives a straight line in the dilute regime:

$$D = D_0[1 + (2MB_2 - \bar{v} - k_s)c] \tag{12.4}$$

In addition to the parameters introduced above, M is the molar mass of the diffusing species, $\bar{v}$ is the partial molar volume of the solvent and k_s is an empirical constant that accounts for the adjustments to the volume fraction that derive from the entrainment of the solvent along the polymer. When plotting D versus c, the slope of the line is equal to $(2B_2 - \bar{v} - k_s)$. For simplicity, we can define this term as;

$$k_D = 2MB_2 - \bar{v} - k_s \tag{12.5}$$

We make two key assumptions in using the equation of Harding and Johnson. First, we assume that the NaCl dependence of the observed diffusion coefficients derives mainly from the NaCl dependence of the solvent-mediated interactions between LAF-1 molecules. Accordingly, our definition of kD is altered to;

$$k_D^{[\mathrm{NaCl}]} = 2MB_2^{[\mathrm{NaCl}]} - \bar{v} - k_s \tag{12.6}$$

As noted above, the partial molar volume of the solvent and the empirical constant k_s (in units of mL /mg) quantify the degree of solvent entrainment. By assuming the form for k_D that is shown in equation 12.6 we are stipulating that $\bar{v}$ and k_s are independent of NaCl concentration. We believe this assumption is justified by the recognition that all

our FCS measurements are quite unlike the sedimentation velocity measurements (through which equation 12.4 was originally defined) in that they are made under dilute protein concentrations and away from the sedimentation regime, where the backflow of solvent upon sedimentation creates a problem with interpreting the impact on measured velocities in sedimentation velocity analytical ultracentrifugation experiments. Additionally, experiments show that the self-diffusion coefficient of water decreases only by roughly 4% in the absence of NaCl when compared to the corresponding value in 1 M NaCl. This provides a proxy for estimating the extent of change to the solvent entrainment and suggests that the magnitude of the change we expect to the term will be at least two orders of magnitude smaller than the measured changes to the diffusion coefficient D as a function of protein concentration c (see fig. 12.6).

Our data show that the slope of the plot of D vs. c decreases as the NaCl concentration increases. We interpret this to imply a weakening of the homotypic associations between LAF-1/ RGG molecules as a function of NaCl. In the presence of high NaCl (1 M NaCl), we note that the diffusion coefficient D varies negligibly with protein concentration c. An example of the insensitivity of D to changes in protein concentration at high salt is shown in fig. 12.6b. Based on these data we make the following second assumption;

$$\bar{v} - k_s = 0 \tag{12.7}$$

This assumption allows us to re-write equation 12.4 as

$$D = D_0[1 + (2MB_{2,app}^{[\mathrm{NaCl}]}c] \tag{12.8}$$

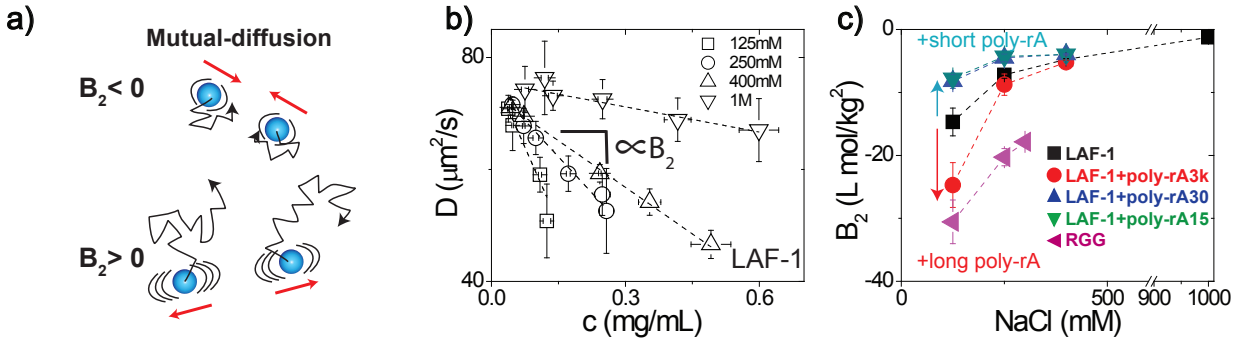

Figure 12.6: RNA and NaCl influence intermolecular interactions of LAF-1 and RGG. (a) Schematic illustration of mutual-diffusion. (b) Mutual diffusion coefficient of LAF-1 strongly depends on the protein concentration. The dashed lines show the linear fits obtained using the equation shown for D in the text. The slope, which is proportional to B_2, decreases with increasing NaCl concentration. (c) The second virial coefficients, B_2, approach the ideal solution limit of zero with increasing NaCl concentration. The B_2 values are most negative across the entire NaCl range for the RGG domain implying stronger effective pairwise interactions for the RGG domain when compared to LAF-1 in the absence or presence of RNA.

We believe that our assumptions are justified by our data and the magnitudes of the changes we observe to the measured diffusion coefficients as a function of protein concentration in different amounts of NaCl. In the equation above, our assumption involves the replacement of the actual second virial coefficient with a NaCl dependent apparent second virial coefficient, $B_{2,app}^{[\mathrm{NaCl}]}$.

In the presence of 1 M NaCl, LAF-1 diffusivity is only moderately dependent on protein concentration, with a slope that is near zero, indicating that LAF-1 is only weakly self-associative at high NaCl concentrations (fig. 12.6b and c). However, as NaCl concentration decreases, LAF-1 diffusivity becomes more strongly dependent on protein concentration, thus

yielding negative B_2 values of increasing magnitude. The addition of short unlabelled RNA molecules gives rise to less negative B_2 values, implying that the RNA molecules weaken the strengths of two-body interactions between LAF-1 molecules. In contrast, in the presence of long RNA molecules we obtain more negative B_2 values, implying a strengthening of the effective two-body interactions between LAF-1 molecules. These results are qualitatively consistent with the changes in diffusivities within the droplets (fig. 12.7). Additionally, measurements of B_2 values for the RGG domain show that these values are the most negative of all the constructs we tested, consistent with this being a highly 'sticky' domain that drives phase separation.

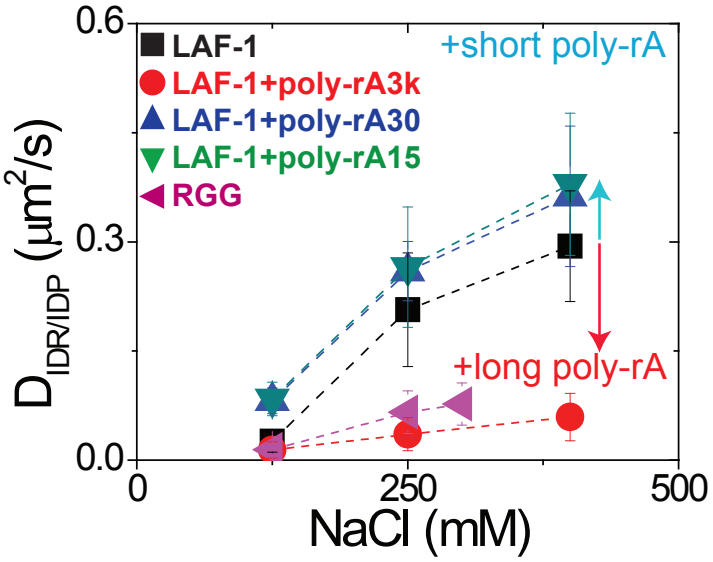

Figure 12.7: Increasing NaCl concentration increases the diffusion coefficients of LAF-1 in droplets. Adding short RNA (poly-rA30 and poly-rA15) also increases diffusion coefficient of LAF-1 in droplets whereas adding long RNA (poly-rA3k) decreases the diffusion coefficients of LAF-1 in droplets.

12.2.4 Quantifying B_2 by 90°Laser Light Scattering

Although we believe the assumptions used to determine the B_2 values are justified, they rely on a number of compounding simplifications, where the impact of error propagation is unclear. To provide an orthogonal test of the assumptions made to extract B_2 values from usFCS measurements, we compared the B_2 values derived from the usFCS measurements with values obtained from right-angle laser light scattering. In this approach, one measures the concentration dependence of the scattered light of the solution with the protein as a function of protein concentration and subtracts the contribution from the buffer alone to uncover the second virial coefficient using a so-called Zimm plot (see fig. 12.8).

Here, the light source was a laser with wavelength (λ) of 488 nm with vertical polarization. Since the molecular size of each of the samples used was smaller than $\lambda/20$, no angular dependence for the excess scattered intensity was expected and all light scattering data were recorded at an angle of 90°. The Rayleigh expression describing the intensity of light scattered from a particle in solution is given in equation

$$\frac{Kc}{R} = \frac{1}{M} + 2B_2c \tag{12.9}$$

where K is an optical constant, c is the particle concentration, R is the Rayleigh ratio of scattered to incident light intensity, M is the molecular weight, and B_2 is the second virial coefficient. The optical constant is defined by equation 12.10;

$$K = \frac{4\pi n^2 \left(\frac{dn}{dc}\right)^2}{N_a\lambda^4} \tag{12.10}$$

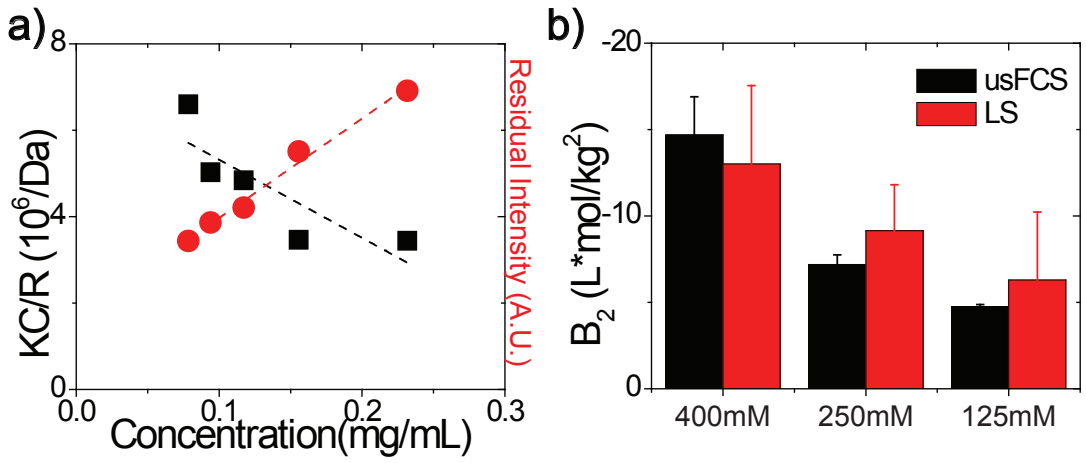

Figure 12.8: (a) Right-angle laser light scattering data for LAF-1 in 400 mM NaCl buffer solution. (b) Comparison of light scattering determined data with the values obtained from usFCS. The estimates of B_2 obtained using both methods are similar within error for three different NaCl concentrations, thus establishing the accuracy of our usFCS measurements and the validity of the assumptions used in our analysis of usFCS data.

where N_A is Avogadro's number, n is the solvent refractive index, and $\frac{dn}{dc}$ is refractive index increment for the protein/solvent ($\sim$0.185 g/mL). The expression used to calculate the sample Rayleigh ratio, R, from a toluene standard is given by equation 12.11

$$R = \frac{I_A n_A^2 R_T}{I_T n_T^2} \tag{12.11}$$

where I_A is the residual scattering intensity of the analyte (sample intensity solvent intensity), I_T is the toluene scattering intensity, n is the solvent refractive index, n_T is the toluene refractive index (1.503 at 488 nm), and R_T is the Rayleigh ratio of toluene (39.6×10^{-6} cm^{-1} at 488 nm).

We compared the measured B_2 values using right angle laser light scattering for LAF-1 at different NaCl concentrations to the values of B_2 that were obtained using usFCS measurements. The values were found to be equivalent across all NaCl concentrations. Therefore, for our analysis in the main text, we used data from usFCS measurements because these afford higher reliability at the low protein concentrations at which these measurements have to be made. Additionally, in contrast to light scattering, the impact of RNA molecules on B_2 can be readily quantified using usFCS measurements because the only signals in these measurements come from labelled molecules.

12.2.5 A Theoretical Framework for the Measured Binodals

The low protein concentration inside LAF-1 droplets comes as a surprise given previous suggestions, including recent measurements of elastin-like-polypeptides (ELPs), which point to concentrations at least two orders of magnitude higher [547]. Moreover, our findings reveal an unusual invariance of critical points and left binodal arms to the addition of RNA, features that we were unable to explained using simple mean-field theories. For example, Flory-Huggins theory suggests that as B_2 becomes more negative (self-interactions are effectively more attractive), c_s should decrease and c_D should increase, but this is not borne out in comparative measurements of the binodals for LAF-1 vs. the RGG domain.

A clue to explaining the curious behaviour comes from all-atom simulations. All-atom Monte Carlo simulations were run using the CAMPARI software package and the ABSINTH implicit solvent model. For further discussion on the ABSINTH forcefield please see chapter 2. We ran 200 short independent simulations to construct an extensive ensemble of ∼100,000 conformations. We also a series of simulations that were ∼10x longer and compared our results

from the 200 short simulations with these longer simulations to assess relative convergence. Both sets of simulations showed statistically identical properties, giving us confidence in the sampling achieved for the ensemble. We quantified the conformational fluctuations of LAF-1 by examining the two-dimensional probability density map of R_G and asphericities. RGG samples a broad range of conformations whereby compact, globular conformations, and expanded coil-like states are sampled with roughly equivalent probabilities (fig. 12.10). The naive expectation is that these conformational features reflect a lack of preference for interaction of chains with themselves vs. solvent, which would be associated with B_2 values of ≈ 0, this is contradicted by our measurements of negative values of B_2 [504]. Therefore, the phase behaviour of RGG domains appears to derive from a combination of large conformational fluctuations, and negative B_2 values; the latter likely arise from polyvalency of 'sticky motifs' comprising charged and aromatic residues. Importantly, the large conformational fluctuations should generate large pervaded volumes, thus dramatically increasing the likelihood that RGG domains will overlap with one another, even at ultra-low concentrations.

The concept of the overlap volume fraction (ϕ^*) is central to describing the phase behavior of polymer solutions. This is the concentration threshold beyond which inter-chain interactions become more likely than intra-chain interactions i.e., the concentration at which different chains begin to overlap significantly with one another. Figure 12.9 provides a graphical representation of the dilute, semidilute, and concentrated regimes.

The overlap concentration threshold defines the boundary between the dilute ($< \phi^*$) and the semidilute ($> \phi^*$) regimes. In a semidilute solution, polymer density fluctuations play a crucial role in determining the interactions between chains [504]. The large conformational fluctuations associated with the RGG domain combined with the low protein density within droplets points to the direct relevance of the physics and chemistry of polymers in semidilute

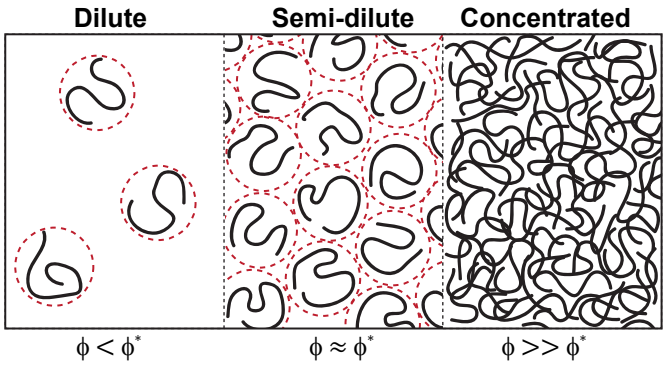

Figure 12.9: Graphical representation of the dilute, semidilute and concentrated regimes. Here, ϕ represents the polymer concentration in the solution and ϕ^* represents the overlap volume fraction. LAF-1 droplets are consistent with a semidilute solution.

solutions. Numerical reproduction of the measured binodals, for both LAF-1 and the RGG domain alone, requires the adaptation of an advanced theory that explicitly accounts for the combined effects of chain density fluctuations, as well as two- and three-body interactions (derived respectively from second and third virial coefficients), as demonstrated in (fig. 12.3b) [409].

Above the overlap concentration, chain density fluctuations will contribute to screening the interactions between pairs of residues on a single chain, provided the spatial distance between residues is larger than the correlation length, ξ. For the semidilute regime this characteristic length scale is also equivalent to the mesh size, as it pertains to the average size of voids between polymer chains. The inferred strengths of three-body interactions are described by w, which for positive values imply a weakening of attractive inter-molecular interactions.

By fitting Muthukumar's theory to the measured binodals, we generate estimates of construct-specific values for ξ and w [409]. Can the derived values obtained from these fits be interpreted as physically useful parameters? ξ provides a measure of the average distance between

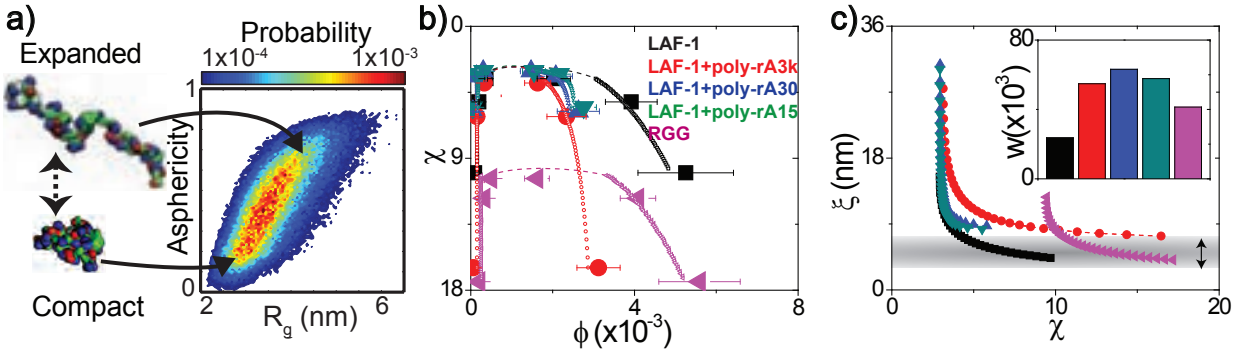

Figure 12.10: Summary of computational and theoretical analysis. (a) Results from atomistic simulations of the RGG domain. (b) Comparison of the binodals derived from numerical implementation of Muthukumar's theory (open symbols) with experimental data (solid symbols). While the data help to identify the critical-region, the precise critical point cannot be reliably located because it is characterized by fluctuations that occur on all length scales. Therefore, the analysis was restricted to reproducing the low and high concentration arms of the binodals away from the critical point. The dashed lines are drawn to guide the eye. (c) χ-dependent values of ξ are shown for each of the constructs and are calculated as described in chapter 13. The horizontal gray stripe corresponds to values of ξ obtained at 125 mM NaCl. The inset shows inferred values of construct-specific three-body interaction coefficients, w and the color-coding of the bars follows the format used in panels (b) and (c).

polymer chains, or of the lengthscale over which fluctuations occur. As χ decreases (becomes less attractive) ξ grows (fig. 12.10c), initially slowly, but as the critical point is approached ξ grows asymptotically towards ∞. In the statistical mechanics of phase transitions the critical point represents the a state at which fluctuations occur over all lengthscales, typically a description cited as a theoretical limit. In this case we see a real, physical manifestation of this approach to infinity through the analytically solved ξ values. At lower NaCl values the correlation length plateaus to a value that appears to depend on the presence or absence

of RNA. In the absence of RNA ξ approaches around 4 nm, while in the presence of RNA (both long and short RNA) this value is larger ($\sim$7 nm). We interpret this to mean that RNA increases the average spacing between LAF-1 molecules, possibly through the intrinsic RNA-RNA repulsion and/or it's increased steric bulk. The three-body interaction parameter w can be interpreted as the excluded volume occupied by the chain. The addition of RNA increases the apparent excluded volume of LAF-1, effectively diluting the concentration of LAF-1 (a result that makes sense given the increase in the correlation length). Surprisingly the excluded volume of LAF-1 alone is lower than that of the RGG domain. One possible interpretation of this is that the RGG domain interacts with the remainder of the protein, effectively prepaying the entropy cost associated with increased compaction via a high effective concentration of helicase domain and the second IDR.

The overlap concentration (ϕ^*) can be directly calculated if the polymer dimensions are known. We calculated the overlap concentration for the RGG domain using the dimensions obtained from simulation, obtaining a value of 8.8×10^{-3}. The predicted ϕ^* is of the same magnitude as the measured c_D. This implies that ξ should be quantitatively similar to the dimensions of an individual molecule. Our numerical reproduction of the measured binodals yields estimates for ξ as a function of χ (fig. 12.10c). At 125 mM NaCl, the predicted value of ξ is between 3 and 8 nm. This range quantitatively matches the average dimensions of the RGG domain inferred from simulations (between 3 and 5 nm, see fig. 12.10); the agreement of these two sets of estimates is remarkable given that they were determined through entirely independent approaches. Our analysis reveals that the protein concentrations in the dense phase is of the same magnitude as the very low overlap concentrations (ϕ^*), which arise from large-scale conformational fluctuations of individual molecules. Therefore, chemical information in the form of sequence-encoded conformational fluctuations controls the dimensions of the disordered RGG domain, and the resulting droplet phase behavior.

12.2.6 Droplet Nanorheology

To quantify the impact of low intra-droplet concentration/volume fraction on molecular motions and rheological properties of droplets, we used usFCS to determine the diffusion coefficients of embedded 14 nm fluorescent spherical nanoparticles. We then use the Stokes-Einstein relation to calculate the viscosity.

$$\eta = \frac{k_B T}{6\pi R D} \tag{12.12}$$

Here, $k_B T$ is the thermal energy scale, R is the nanoparticle radius, and D is the measured diffusion coefficient. For LAF-1 in 125 mM NaCl, this analysis reveals droplet viscosity of 27.2 ± 5.9 Pa·s, a value consistent with measurements based on particle tracking microrheology. Using a similar approach, we found that RGG droplets are roughly twice as viscous as full-length LAF-1 droplets (fig. 12.11a), in agreement with the finding that B_2 values for RGG are significantly more negative than full-length LAF-1. Moreover, measurements within RGG droplets show that the RGG domains diffuse more slowly in these droplets when compared to full-length LAF-1 molecules in droplets formed by full-length LAF-1 (fig. 12.7). Both full-length LAF-1 and RGG droplets exhibit a decreased viscosity and increased molecular diffusivity upon increasing NaCl concentration (fig. 12.7 and fig. 12.11a). This is consistent with the decreasing magnitudes of B_2 values as NaCl concentrations increase (fig. 12.6c).

When we added RNA of either 15 or 30 nucleotides into LAF-1 droplets *in vitro*, the droplet viscosity decreased to 16.1 ± 2.8 Pa·s at 125 mM NaCl. A similar mass concentration of longer (3,000 nucleotide) poly-rA caused the opposite effect, whereby the droplet viscosity increases to 60.9 ± 10.3 Pa·s. Nonetheless, in all cases, the droplet viscosity still decreases as the concentration of NaCl was increased. These changes are also mirrored in the diffusivities

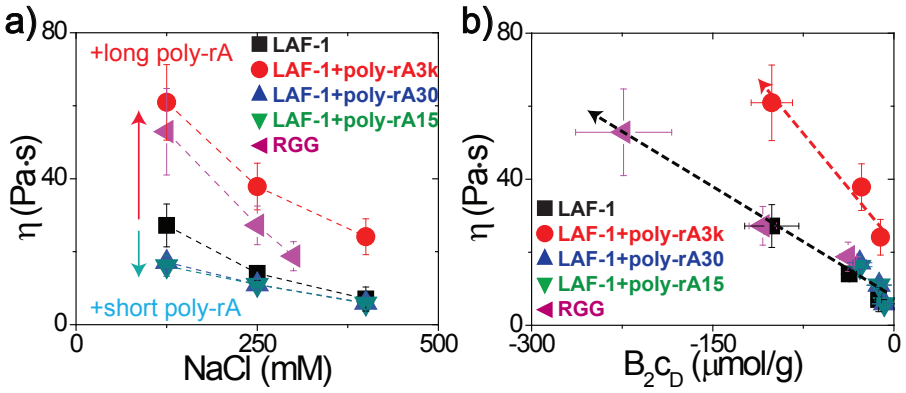

Figure 12.11: Nano-scale rheology of RGG and LAF-1 condensed droplets. (a) Increasing the concentration of NaCl decreases the viscosity within LAF-1 droplets. Adding short RNA (poly-rA30 and poly-rA15) also decreases the viscosity within LAF-1 droplets. However, adding long RNA (poly-rA3k) increases viscosity of LAF-1 droplets. (b) Viscosity within droplets is proportional to the product of B_2 c_D. Upon addition of the short RNA (poly-rA) the droplet viscosity decreases and follows the same universal curve with LAF-1 and RGG.

of molecules within the droplets (fig. 12.7). However, changes in droplet viscosity are not fully captured by considering B_2 alone. Consistent with simple theories of viscosity in polymeric systems, changes in droplet viscosity also depend on the protein concentration within the droplet, c_D ; these combined effects can be captured by plotting viscosity as a product of, $B_2 c_D$, as shown in fig. 12.11b. Interestingly, however, the ability to collapse viscosity data as a function of data $B_2 c_D$ breaks down for the long poly-rA (3,000). We speculate that this may reflect the fact that there is a many:1 stochiometry between LAF-1 and the longer RNA, suggesting the longer RNA may allow the droplet interior to percolate, fundamentally altering the internal dynamics when compared to droplets where the RNA does not provide this connectivity.

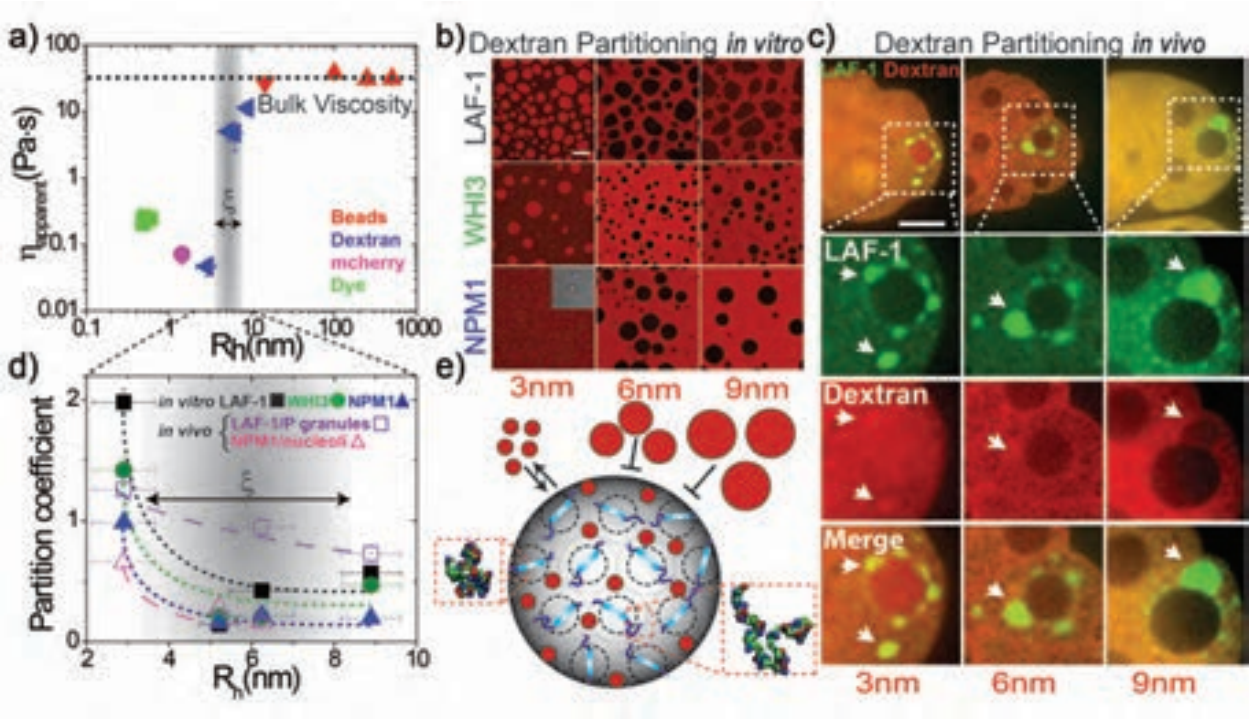

Figure 12.12: Low-density semidilute liquid droplets. (a) Apparent viscosities extracted from measurements of diffusion coefficients of probes within LAF-1 droplets at 125 mM NaCl. The gray bar corresponds to ξ in fig. 12.10c (b) Permeability of different *in vitro* droplets to fluorescent dextran (red). The inset figure shows the bright-field image of NPM1 droplets. (c) Permeability of *in vivo* LAF-1::GFP labelled P granules in *C. elegans* to fluorescent dextran. Perinuclear P granules in ∼16-cell embryos are indicated with arrows. (Scale bar, 10 µm). (d) Partition coefficients were calculated from fluorescent intensities inside/outside droplets. The gray bar corresponds to ξ in fig. 12.12a and 12.10c. (e) Schematic model. The RGG domain is depicted in blue and the envelope is defined by the R_G of LAF-1 are shown in black-dash circles.

The agreement between the viscosity determined from the diffusive motion of 14 nm particles and micron-sized particles (fig. 12.12) suggests that the effective mesh size (ξ) of the intra-droplet protein network is less than 14 nm. To infer the value of for LAF-1 droplets, we measured the diffusion coefficient for a variety of molecular probes of smaller sizes, and use their hydrodynamic radius R_H to calculate an apparent viscosity as above; we note that using the Stokes-Einstein equation to estimate viscosity is only strictly valid for spherical probe particles, a point that is discussed in depth in the following section. Small solutes (R_H ~0.5 nm) and the globular protein mCherry (R_H ~1.4 nm) exhibit values in the range of 0.07 - 0.2 Pa·s. These values are roughly two orders of magnitude lower than the bulk viscosity, consistent with their motion primarily reflecting diffusion through the aqueous solvent that permeates the droplet mesh. To interrogate larger length scales, we used dextran molecules of differing molecular weights. In dilute aqueous buffers, the 10 kDa dextran has a R_H of ~2.3 nm. By plugging this value of R_H and the measured diffusion coefficient into the Stokes-Einstein equation, we obtain an apparent viscosity value that is comparable to those of the other small probes. However, a similar analysis applied to the measured diffusion coefficients of 40 kDa (R_H ~4.5 nm) and larger molecular weight dextran molecules suggests significantly hindered motion, implying that the bulk properties of the droplet become increasingly dominant. Our dextran diffusivity data are also consistent with the partitioning of different molecular weight dextran into droplets. We find that 10 kDa dextran molecules strongly partition into LAF-1 droplets (fig. 12.12c, 12.12e), while 70 kDa and 155 kDa dextran molecules (¿ 6 nm) are mostly excluded. These findings suggest that the characteristic mesh size within droplets is between 3 and 6 nm, in agreement with results from our theoretical analysis and simulations (Fig 12.10c).

12.2.7 Nanorheology of Polymer Solutions

Since dextran molecules, especially of higher molecular weights, are not well approximated as spheres, we also analyzed the diffusivity data using the framework of Cai *et al.* [82].

Unlike spherical probes, dextran molecules are flexible polymers. As a result, the dynamics of a polymeric probe (dextran) in a polymeric solution (LAF-1/RGG) may not be well described by a Stokes-Einstein based relationship. For probe molecular weights below a threshold value, probe molecules will behave like small solutes. Conversely, above some threshold molecular weight, the diffusivities of dextran molecules should decrease as the reciprocal of increasing molecular weight. This thresholding behaviour is well described by theoretical analysis by Cai *et al.*, who describe several distinct regimes.

Dextran is frequently described as behaving like an ideal chain in aqueous buffers. To verify this, we analyzed published data for hydrodynamic radii as a function of molecular weight. The measurements shown in based on viscosity analysis and light scattering. Here, we plot data from the literature as a log-log plot, with the log of the R_H along the x-axis and the log of the degree of polymerization (N) along the y-axis. This analysis yields a straight line with a slope that corresponds to the scaling exponent ν and an intercept that corresponds to $\log(b)$, where b is the Kuhn length. We find $\nu = 0.49$, which is in good agreement with the theoretical exponent of 0.5 expected for an ideal chain. The intercept yields a value of b = 0.303 nm. For molecular weights below a threshold value, dextran molecules should behave like small solutes. Conversely, above a threshold molecular weight, denoted as ξ_L,, the diffusivities of Dextran molecules should decrease as the reciprocal of increasing molecular weight.

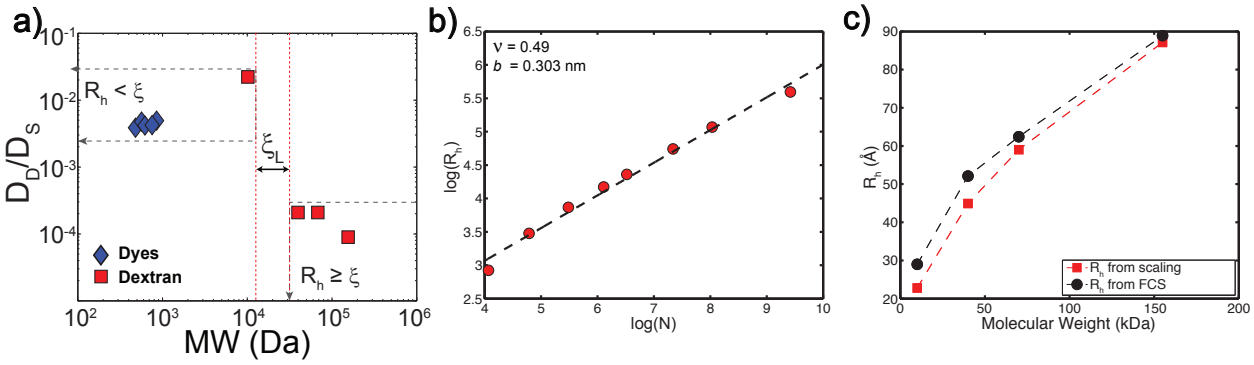

Figure 12.13: (a) Comparison of diffusivity behaviour in two of the regimes, where ξ_L defines the the threshold lengthscale at which the probe size is approximately equal to the mesh-size. (b) Experimentally derived scaling behaviour for dextran demonstrates that dextran behaves as an ideal chain in solvent (in agreement with previously published results). (c) Comparison of R_H values obtained from FCS with those extracted from theory given the derived scaling behaviour ($\nu = 0.5$) and Kuhn length ($b = 0.303$ nm) of dextran

We measured the diffusivities of dextran molecules of four different molecular weights in droplets (D_D) and in bulk solution (D_S). Following the theoretical analysis of Cai *et al.*, we plotted the ratios of D_D to D_S as a function of molecular weight. This analysis shows that the ratio of diffusion coefficients for the 10 kDa dextran molecule is within an order of magnitude of the values for small fluorescent dyes, despite the fact that the molecular weights are different by two orders of magnitude. In contrast, beyond a threshold molecular weight of ∼40 kDa, the ratio of diffusivities for dextran molecules decrease by at least two orders of magnitude and show a decrease with increased molecular weight dependence that is expected of polymeric probes. Based on these results, the dextran molecular weight that corresponds to the lower bound of the mesh size lies between 10 kDa and 40 kDa. The results are in line with theoretical predictions, and point to the existence of a threshold lower bound of the mesh-size, delineated by ξ_L.

A threshold Mw of 40 kDa yields a dextran degree of polymerization of ≈ 245. Using N and the the well defined relationship defined in 12.13b we estimate ξ_L to be $bN^{0.5} \approx 4.7\,\text{nm}$. We use $b = 0.303$ nm, which is the Kuhn length of dextran molecules obtained through the linear fit in 12.13b. Finally, to assess the robustness of our these data we compared the inferred R_H of dextran in solution obtained by the linear fit with the value obtained via usFCS. These two methods yielded results that show good agreement with one another.

12.2.8 Low Density Droplets Persist *in vivo* and Across Different Systems

LAF-1 is a critical component for the formation and stability of P-granules in *C. elegans*. However, despite LAF-1 being necessary for P-granule formation, it is not sufficient. P-granules are complex organelles containing a range of proteins and RNA. Our *in vitro* experiments consistent of a two (or three) component mixture, a far cry from the complex melange associated with P-granules. We wondered if the droplet permeability and mesh-size threshold observed *in vitro* is simply an artifact of our experiments. To challenge this concern directly we measured the partitioning of dextran into LAF-1::GFP labelled P granules in *C. elegans* embryos (fig. 12.12c, 12.12d). Dextrans were diluted to 4 mg/mL in injection buffer (20 mM KPo4 pH 7.5, 3 mM K citrate, 2% peg-1000) were injected into the syncytial gonad of LAF-1-crispr worms. After incubation for 4-5 hours at 25°, worm gonads were dissected and embryos were imaged by confocal microscopy.

Consistent with our *in vitro* data, we find that the smaller 10 kDa dextran partitions favorably into P granules, while the larger 155 kDa dextran is significantly exclude.

For the *in vitro* results, the the 70 kDa dextran ($R_H \approx 6$ nm) was excluded from the droplets (fig. 12.12b), while *in vivo* this same dextran was able to enter (fig. 12.12c). P-granules contain both protein and RNA and our analysis of the *in vitro* phase diagrams suggests that RNA leads to an increase in the meshsize (12.10. Taken together, the difference between the *in vitro* and *in vivo* results are an entirely expected difference that further suggests our theoretical analysis provides genuine insight into the material properties of the phase separated state.

Finally, we sought to assess if these results were specific to LAF-1, or provided more general insight into membraneless organelles. We see similar size-dependent exclusion *in vitro* for two other well-known intrinsically disordered droplet forming proteins, WHI3 and NPM1 (fig. 12.12b and d). These results suggest that the semidilute, void-rich nature of droplets is may be a feature of many liquid phase organelles that are driven by the sequence-encoded conformational fluctuations of IDPs/IDRs. We emphasize that this does not preclude the formation of dense IDR mediated droplets, but tentatively suggest that void-filled droplets may be a necessary feature of membraneless organelles, distinct phases droplets that have evolved to accommodate and internalize client species.

12.3 Discussion

Phase separation has been recognized as a ubiquitous mechanism for organizing the contents of living cells. IDRs of proteins appear to play an important role in driving phase separation. However, there is a lack of clarity regarding the connection between the sequence-encoded conformational fluctuations of heterogeneous molecules and the microscale organization and dynamics of droplets that results from phase separation. Here, we begin to uncover these

connections using a new usFCS approach to measure the phase behavior and intra-droplet properties of droplets formed by LAF-1, which contains a disordered R/G-rich domain that is necessary and sufficient to drive phase separation. By quantifying the strengths of pairwise interactions and resulting coexistence curves, together with measurements of nanoscale viscosity, molecular partitioning, and theoretical analysis, our results provide a holistic picture of how emergent properties of membraneless organelles derive from the amplitudes of conformational fluctuations and interactions of component molecules.

As in many membraneless organelles, LAF-1 and other P granule proteins function by interacting with RNA [601]. Previous work had shown that short (50 nt) RNA molecules decrease the viscosity of LAF-1 droplets, consistent with RNA impacting protein-protein interactions. However, in that work, RNA had no effect on the saturation concentration required for phase separation, an unusual result given the coexistence curve should arise from the same molecular interactions that govern droplet dynamics. Our results here help resolve this apparent paradox. We find that RNA does indeed impact the coexistence curve, albeit by shifting only the high concentration arm to lower values, while leaving the low concentration arm and critical point invariant. Interestingly, while our results reveal a robust shift in the presence of RNA for all lengths tested (15 - 3000 nt), we find that only short RNA decreases droplet viscosity, while long RNA has the opposite effect. This suggests that additional physical processes such as protein-RNA entanglements might be important in describing molecular transport in the presence of longer RNA molecules. Systematic investigations of ternary phase diagrams are needed to obtain a complete understanding of how polydispersity of RNA lengths, sequences, and structural motifs regulate the overall phase behavior of proteins such as LAF-1. This is biologically relevant because RNA molecules of varying lengths are found in P granules. We speculate that their relative abundance could

tune P granule viscosity and phase behavior by modulating the effective interactions between LAF-1 molecules or entangling with them.

Our results have implications for the interpretation of phase diagrams derived from components that undergo biological phase separation. The changes to our LAF-1 phase diagrams are asymmetric, with the low and high concentration binodals changing to different extents. While RNA has no impact on the low concentration arm, it reduces the width of the two phase regime by shifting the high concentration arm of the binodal, a result that can be interpreted in terms of an increase in the mesh-size, an emergent impact that alters that condensed phase without changing the dilute phase properties. There are many other examples of RNA, osmolytes, and other proteins altering the phase behaviour of proteins that drive biological phase separation. These systems have so far only been studied in terms of the impact of clients on the low concentration of the binodal. Our results illustrate that an invariant low concentration binodal cannot *necessarily* be interpreted as insensitivity to a client protein. Similarly, clients that do alter the low concentration arms may alter the high concentration arms to a greater or lesser extent. As an example, in original work on LAF-1 the deletion of the C-terminal IDR was found to have no impact on the low concentration arm of the binodal. However, the high concentration of the binodal was not measured at this time, raising the possibility that the C-terminal IDR influences the high concentration arm but not the low concentration arm.

A surprising finding is that LAF-1 and two other proteins with IDRs phase separate into liquid droplets of ultra-low protein concentrations that correspond to the semidilute regime. We identify the characteristic mesh size within these permeable droplets to be $\sim$3 - 8 nm. To account for this behavior, we adapted an analytical model that explicitly accounts for the effects of conformational and chain density fluctuations. For LAF-1, the key region

for understanding the multiscale structural features of droplets is the RGG domain, which is necessary and sufficient for phase separation, and underlies the intriguing properties of P-granule-like LAF-1 droplets. In particular, the sequence of the RGG domain imparts a unique and unexpected combination of attractive interactions (strongly negative B_2), with large-scale conformational fluctuations and average preference for expanded conformations. These combined effects readily engender overlap among chain molecules, even at very low protein concentrations. This allows the RGG domain to drive the LAF-1:LAF-1 interactions underlying phase separation, while still resulting in remarkably low-density droplets.

Another archetypal IDP whose phase behavior adheres to the predictions of mean field theories has been described in a recent study that characterized the full phase behavior of elastin-like polypeptides (ELPs) [547]. These IDPs are bereft of charged residues, and in dilute solutions above an upper critical solution temperature (UCST) are predicted to form compact globules. Above a temperature-dependent saturation concentration ELPs coalesce and entangle to drive phase separation. The measured binodals reveal concentrations for c_S and c_D that are at least two orders of magnitude larger than those measured here for LAF-1 and the RGG domain. These differences can be rationalized in terms of the sequence-encoded globule versus coil-like behavior of ELPs versus RGG domains and the differences in the amplitudes of conformational fluctuations. Taken together, the recent study of Simon *et al.* and our results highlight the direct connections between sequence-encoded and context dependent conformational fluctuations of IDPs and the driving forces for phase separation that govern the final, salt / temperature dependent values of c_S and c_D. A key challenge for future work is to uncover the specific molecular-level driving forces of liquid-liquid phase separation for complex IDPs that are known encode a range of different sequence-to-conformation relationships in different environments.

Our results reveal that LAF-1 droplets, as well as intracellular RNA/protein droplets, are dense when compared to the surrounding solution, but are nevertheless solvent-rich and full of permeable voids that accommodate the free diffusion of small solutes, folded proteins, and flexible polymers up to a specific threshold in size/ molecular weight. We find that molecular scale motions within droplets can be decoupled from the mesoscale droplet properties. For example, the bulk viscosity of the droplet has little affect on the diffusion of molecules that are smaller than the mesh size, because they are free to move through the free volume within the droplet. In contrast, larger macromolecules and complexes recruited within droplets will be subject to the viscous drag arising from the dynamic network of IDR-self associations. Since protein sizes span the droplet mesh size, from around 2 nm for an average monomeric protein, to $>$ 10 nm for multimeric complexes, these effects are likely to have significant consequences for intracellular droplet functions. We further speculate that low density droplets with large mesh-sizes could allow for size-selective filtering, which could potentially be regulated, in a manner analogous to that of FG Nups in the nuclear pore, which exhibit a comparable passive mesh size ($\sim$4 nm).

Our findings are likely to shed light not only on P granules, but also many other membraneless organelles. Indeed, IDRs harboring R/G-rich sequences similar to that of LAF-1 are found in many RNA binding proteins, including those that are known to drive phase separation7, 11, 20, 44. Our findings may thus provide a new framework for understanding the length scale dependent properties of low-density liquid phase organelles throughout the cell. Changes in these properties will be relevant not only for physiological function, but also in disease-associated pathological aggregation.

Chapter 13

Numerical and Analytical Approaches for Fitting Phase Diagrams

The work in this chapter is in preparation for a forthcoming manuscript. Any experimental work discussed in this chapter was performed by Steven Wei and Shani Elbaum-Garfinkle. All the described theoretical and simulation work was done by A.S.H and R.V.P.

13.1 Background and Motivation

Over the last ten years biological phase separation has emerged as a critical mechanism for cellular organization [27, 251, 391]. The weakly-attractive multivalent interactions that in many cases appear to drive the formation of liquid-like biomolecular condensates (also known as liquid droplets, membraneless organelles, liquid-like assemblies, quinary assemblies, and various other names) allow chemically specific compartments to dynamically assemble and disassemble in a highly regulatable yet efficient manner. Other condensates show more solid-like characteristics, and suggest a coupling between phase separation and gelation [483, 639].

In chapter 3 we introduce distinctions (semantic or otherwise) between various types of assemblies, but for the purposes of this chapter we are only referring to large, micron-scale, liquid-like intracellular assemblies. As a result, we will use the terms condensate, droplet, and liquid-like assemblies interchangeably.

Why might intracellular droplets form? They provide a means for the cells to specifically and robustly assemble specific components into well defined compartments. The selectivity associated with these liquid-like droplets is poorly understood, but it likely represents an emergent property of the droplet's material properties and chemical composition. In the case of proteins, this selectivity likely originate from a combination of the protein's size, and the the composition and distribution of amino acids [67, 420]. In the case of RNA, these preferences are likely a combination of secondary and tertiary structure and nucleotide sequence [668]. In many (though not all) cases, the proteins that drive the formation of these condensates contain intrinsically disordered regions. Unlike folded proteins, these regions are unable to autonomously fold into well-defined three dimensional structures, and instead exist in an ensemble of states [603,608]. Given that folded proteins can undergo liquid-liquid phase separation via interaction domains connected by flexible linkers, it had been unclear as to why intrinsically disordered proteins should so frequently found associated with the key drives of biological phase separation [330].

The first organelles shown to behave as a phase separated liquid were P-granules [65]. LAF-1 represents one of the best studied proteins necessary for P-granule formation - *in vitro* LAF-1 forms liquid-like droplets, making it a convenient test system [162]. The disordered RGG domain of LAF-1 is necessary and sufficient to drive LAF-1 droplet formation [162]. In chapter 12, we demonstrated that the RGG domain simultaneously engages in large-scale conformational fluctuations, yet has a strongly negative B_2. The combination of these two

features gives rise to a hetero-polymer that readily forms low-density protein droplets. This provides a directly link between the conformational behaviour associated with disordered proteins and the material properties of a liquid droplet.

Phase diagrams provide a convenient analytical framework to quantitatively describe the phase behaviour of a solution. Figure 13.1 shows phase diagram schematic, where the abscissa describes protein concentration and the ordinate reflects some physical parameter which typically determines the strength of interaction between monomeric units. In principle a phase diagram is a plot that describes the state of two conditions plotted against one another, and defines the at equilibrium thermodynamic phase experienced by the system at a given intersection of those two states [140]. In the world of biological phase separation these phase diagrams have traditionally be drawn as concentration-salt or concentration-temperature phase diagrams [79, 162, 399, 421]. Concentration should ideally be described in terms of mass concentration or volume fraction molar concentration has the unfortunate side-effect of removing any information relating to the molecular nature of the constituent monomer and can be highly misleading when comparing the phase behaviour of proteins with different molecular mass.

The coexistence curve that delinates the two-phase region from the one-phase region is referred to as the **binodal**. The binodal envelopes a second, similarly shaped curve referred to as the **spinodal**. The space between the spinodal and the binodal is metastable the two phase (de-mixed) state is the true thermodynamic minimum, but for a solution in this metastable region, phase separation will proceed via a **nucleation dependent** mechanism. This metastability is because the solution is stable with respect to the natural compositional fluctuations that occur at the finite temperature. In contrast, for a solution where conditions lie inside the spinodal, any kind of fluctuation will lead to **spinodal decomposition** and

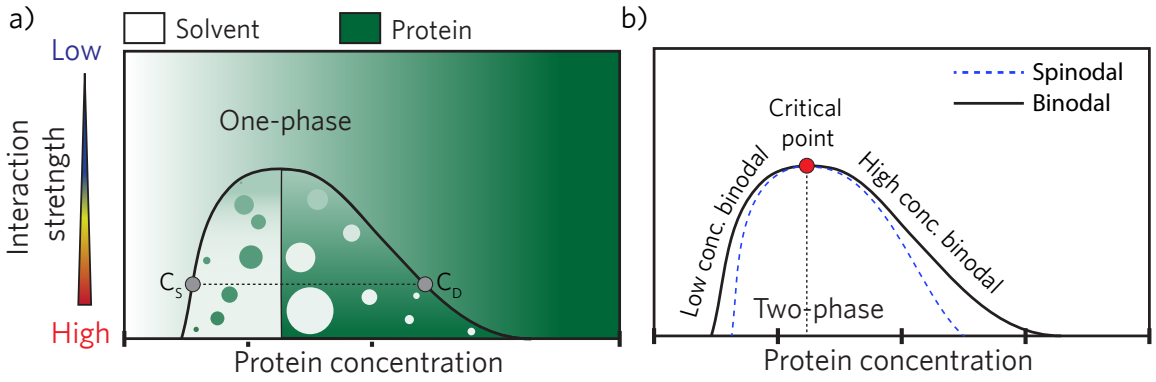

Figure 13.1: A schematic of a phase diagram. Both panels describe the same phase diagram but highlight different features discussed in the text.

the formation of a two-phase equilibria [504]. The binodal and spinodals intersect at the **critical point**.

For the work in this chapter we will describe phase diagrams drawn either as salt vs. volume fraction, or interaction strength vs. volume fraction. For LAF-1 and the RGG domain, NaCl weakens the intermolecular interactions such that with an appropriate conversion mechanism these two descriptions are equivalent. Although salt has been shown to weaken the intermolecular interactions in several disordered proteins that undergo phase separation, this does not necessarily have to be the case [399, 421]. For example, the disordered protein FUS undergoes phase separation more readily at increasing salt concentration [79]. Considering this, we caution that although varying the salt concentration provides a route to tune the interaction strength in the LAF-1 system, this may not necessarily be the case for other systems that undergo biological phase separation.

For a system in the two-phase regime (i.e. at a solute concentration and interaction strength such that the system exists inside the binodal curves), the concentrations associated with the dilute and dense phases are fixed, irrespective of *total* protein concentration. As additional

proteins is added to the system, the volume of the dense-phase increases and, correspondingly, the volume of the dilute phase decreases. In this way, phase separation can be thought of as as having a buffering effect on the bulk concentration of the solute of interest.

In our work on LAF-1 we used a novel ultra-fast scanning fluorescence correlation spectroscopy (usFCS) method to directly measure the concentration in the dilute and dense phases (see chapter 12 for experimental details). These results provided the first examples of full binodals curves of a disordered protein, and to our surprise demonstrated that the concentration within LAF-1 droplets is only $\sim 50\times$ higher than the dilute phase, in the 4-7 mg/ml region. For context, the expected/hypothesized concentration for a disordered protein is in the $\sim$ 100-300 mg/ml [79, 337]. Moreover, the intra-droplet concentration of liquid-liquid phase separated lysozyme has been measured at $\sim$600 mg/ml, and droplets of disordered elastin-like polypeptides were found to be around 300 mg/ml [69, 547]. In these previously studied systems, the phase behaviour is well described by standard Flory-Huggins style theories (we note that the work by Lin *et al.* introduces a novel random phase approximation approach to directly capture the sequence patterning, but in spirit remains a mean-field description). In contrast, the data collected for LAF-1 cannot be fit by Flory-Huggins theory. Instead, we were only able to fit the measured binodals using an advanced mixing theory that directly takes the conformational fluctuations of the underlying polymer into account [409, 410]. Taken together, these results suggest that the underlying phase behaviour of LAF-1 is fundamentally different from many proteins that undergo phase separation.

The remainder of this chapter is laid out as follows. Firstly, we will provide a relatively complete overview of the thermodynamics of mixing in polymer systems. This includes a full derivation of the Flory-Huggins free energy of mixing from first principles, primarily

because this introduces many of the ideas critical to understanding how the free energy mixing relates to the binodal curves. Secondly, we will review the details associated with Muthukumar's theory of polymer solutions. Thirdly, we will discuss the practical challenges associated with calculating extracting the binodals and spiodals from a free energy of mixing expression. Finally (and of perhaps most interest to the majority of readers) we will provide a general, high-level description of the underlying physics associated with the formation of these dilute droplets. We will then outline a general hypothesis that there are (at an absolute minimum) two distinct types of biological phase separation. One is characterized by the larger dilute droplets that one typically associates with membraneless organelles. The second is characterized by much smaller and much denser droplets responsible for localization and sequestration. The reality is these are likely two ends of a continuous distribution of droplet densities found throughout Nature.

13.2 Thermodynamics of Polymer Mixing

The following section introduces the key concepts associated with polymer mixing theories, deriving the energy and entropy of mixing from first principles with an intuitive description of what these terms are describing in the Flory-Huggins theory framework [179–181,247,504]. We will not derive Muthukumar's theory with anywhere near the same detail as we do for the Flory-Huggins free energy of mixing, but introducing the key concepts will be important for understanding how we are able to construct phase diagrams from free energy of mixing curves, as described in a later section.

13.2.1 Polymer Volume Fraction

In the interest of clarity, before we begin we will define several terms used frequently in polymer chemistry which may have subtly different meanings depending on the context. A **monomer** is a single subunit. A **polymer** is the molecule made up of many monomers connected head-to-toe, and is also frequently described as a single chain. The **degree of polymerization** (noted as r) refers to the number of monomers in a polymer. Finally, the **correlation length** that we describe in this work refers to fluctuations in chain density/protein concentration, and can equivalently be thought of as the screening length, assuming we are in the semidilute regime (as discussed later). We note that the definition of the correlation length is context dependent, but in general within the polymer literature ξ is used to denote this the concentration-fluctuation defined correlation length.

Let us consider a lattice - a three dimensional grid of positions evenly spaced apart. Each position on the lattice is referred to as a site, and each site may be occupied by a black

sphere (the solute), or it may be empty (and be occupied by solvent). The total volume of the lattice is given by

$$V_{total} = V_A + V_0 \tag{13.1}$$

Here, V_A is the total volume of the lattice occupied by black spheres, while V_0 is the total volume of the lattice occupied by empty space. For a simple (non-polymeric) solute, each 'molecule' of solution occupies a single site on the lattice; consequently $V_A = N_A$, where N_A reflects the number of molecules of type A in the system. For a polymer (where each molecule of A consists of two or more spheres), $V_A = N_A r$, where r is the degree of polymerization - the number of monomers in an individual polymer. For polypeptides, r is typically fixed; all our molecules of the RGG domain have exactly 168 amino acids. In synthetic chemistry, polymer solutions are often poly-disperse, meaning there is a distribution of r around some mean value. This introduces various complications which we will not expand upon here.

A convenient metric for thinking about polymer solutions is that of volume fraction (ϕ). ϕ refers to the fraction of the lattice occupied by a polymer. In our example above, the volume fraction of occupied sites is given by

$$\phi_A = \frac{V_A}{V_A + V_0} \tag{13.2}$$

and the volume fraction of empty sites is given by

$$\phi_0 = \frac{V_0}{V_A + V_0} = 1 - \phi_A \tag{13.3}$$

Finally, the total number of sites in the lattice is n. If each monomer occupies a single site on the lattice then $n\phi_A = V_A$. This need not necessarily be the case, but for the purposes of our discussion here will will assume this to be true. For a graphical representation of how these parameters relate to a real lattice see fig. 13.2.1.

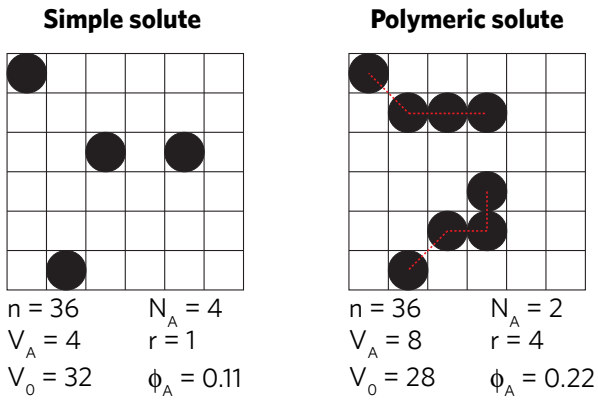

Figure 13.2: A schematic showing how the various parameters defined in this section relate to a simple solute (black spheres) in a two dimensional lattice. The red-dashed lines in the polymeric solute panel illustrates bonds between beads.

It is crucial to choose the correct set of units for quantifying concentrations in polymer solutions. Quantities such as molar concentrations or mole fractions are problematic; they place the high molecular weight polymers and low molecular weight solvent molecules on an equal footing, representing concentration as 'numbers of copies of X'. This yields misleading inferences regarding the phase behaviour of polymer solutions. Consequently, the use of volume fraction is critical. Polymers solutions are classified as being dilute, semidilute, or concentrated. This classification depends on the polymer mass concentration (c_A) typically measured in units of mg / mL or the volume fraction. These two measures of concentration are related through the intrinsic polymer density ρ_0

$$c_A = \phi_A \rho_0 \tag{13.4}$$

Here, ρ_0 is a conversion factor that describes the solution density if the entire solution were solute. For protein systems, we estimate this conversion factor to be ∼1310.16 mg/mL, although the degree of precision here is perhaps misleading. This number originates from the approximate volume of a single amino acid being 140 Å^3 and an average mass of 110 mg/ml The exact value associated with ρ_0 has a relatively low impact on the derived volume fraction value. For example, in a solution where $c_A = 10$ mg/ml, ϕ_A is expected to be 0.0076, (0.76%), although for $\rho_0 \pm 100$ we obtain $\phi_A = 0.0071$ or $\phi_A = 0.0083$.

13.2.2 The Entropy of Mixing

What are the rules that govern the mixing of a two-component system? There are two factors that influence the mixing/demixing of multi-component solutions: the strength of the interactions between the constituent components, and the entropy associated with mixing. The entropy of a system (S) is simply the natural logarithm of the number of different configurations of that system (Ω) multiplied by Boltzmann's constant (k_B) (see 13.5). Abstractly, we can think of Boltzmann's constant as a conversion factor.

$$S = k_B \ln(\Omega) \tag{13.5}$$

In a one-component system (e.g. $\phi_A = 1$) the number of possible states for a single molecule is $\Omega_A = n$. Every site is accessible to the molecule, giving n possible locations across the

lattice that any molecule can occupy. For a fully mixed two-component system consistent of A and 0, the same result holds - $\Omega_{A\text{mixed}} = n$). Again, this is because from the perspective of a given molecule, any site on the lattice is equally accessible. However, for a demixed (phase separated) system the number of possible states for molecule A is given by equation 13.6

$$\Omega_{A,\text{demixed}} = n\phi_A \tag{13.6}$$

Why should this be? In the demixed system, all of the black spheres (A molecules) have assembled together into their own phase. Consequently the accessible volume for any given molecule is no longer *any* site on the lattice, but only those sites included in this A-rich phase, which corresponds to $n\phi_A$ possible sites. We can synthesize these ideas into a single expression to describe the entropy of mixing;

$$\begin{aligned}
\Delta S_{A,\text{demix} \rightarrow \text{ mix}} &= k\ln(\Omega_{A,\text{ mixed}}) - k\ln(\Omega_{A,\text{ demixed}}) \\
&= k\ln(n) - k\ln(n\phi_A) \\
&= k\ln\left(\frac{n}{n\phi_A}\right) \\
&= k\ln\left(\frac{1}{\phi_A}\right) \\
&= -k\ln(\phi_A)
\end{aligned} \tag{13.7}$$

To determine the total entropy of mixing (S_M) we must sum the contributions of all molecules in the system, and normalize by the number of lattice sites, giving the following definition of the entropy of mixing per lattice site ($\Delta\bar{S}_M$);

$$\Delta \bar{S}_M = -k \left(\left(\frac{\phi_A}{N_A} \right) \ln(\phi_A) + \left(\frac{\phi_0}{N_0} \right) \ln(\phi_0) \right) \tag{13.8}$$

The expression above is a general expression for an binary system, but for a binary system with polymer of volume fraction ϕ_A of degree r in solution of volume fraction ϕ_0 (and a degree of polymerization of 1) we can further simplify equation 13.8 to;

$$\Delta \bar{S}_M = -k \left[\left(\frac{\phi_A}{r} \right) \ln(\phi_A) + \phi_0 \ln(\phi_0) \right] \tag{13.9}$$

Note that ϕ_A and ϕ_0 will always lie between 0 and 1, meaning $\Delta \bar{S}_M$ will always be positive. In other words, for simple polymeric models such as our lattice model, the entropy change upon mixing is *always* favourable. As a result, a demixing process must 'fight' the entropic driving force for mixing by providing a sufficient energy of demixing ($\Delta \bar{U}_M$) such that the Helmholtz free energy of mixing ($\Delta \bar{F}_M$) is negative. $\Delta \bar{F}_M$ can be written as

$$\Delta \bar{F}_M = \Delta \bar{U}_M - T \Delta \bar{S}_M \tag{13.10}$$

For demixing to occur $\Delta \bar{U}_M - T \Delta \bar{S}_M \leq 0$, meaning for ideal mixtures (where $\Delta \bar{U}_M = 0$) the thermodynamic minimum is one of a fully mixed solution. This is a somewhat protracted explanation for what is a relatively intuitive result. If we have a room full of people wondering around without a care in the world, they won't all bunch up into one corner (unless that corner has cake), but will aimlessly shuffle around without any strong preferences, leading to a uniform distribution of people across the room. This isn't magic, it's entropy! Now, let us imagine a high-school dance; behold, a demixing phenomena has emerged! Rather than

mixing, the girls and boys awkwardly stand across the room from one another. Here, teenage awkwardness acts as the energy of demixing, which in biomolecular systems is typically not the relevant driving force. Considering this, what *does* facilitates demixing in biomolecular systems, and can we describe it in a correct and intuitive way?

13.2.3 The Energy of Mixing

For Flory-Huggins theory, the energy of mixing is defined by the Flory χ parameter [179,247]. As we will see, while χ is a useful parameter, additional contributions to the energy of mixing can provide a more advanced description of a polymeric system. Never-the-less, we will begin with a brief dissection on the origins of χ.

Let us return to our lattice. The interaction between beads on the lattice may be repulsive, neutral, or attractive. If we assume that two species cannot occupy the same position on the lattice, we are implicitly stipulating a repulsive interaction between sites at the extremely close approach (an excluded volume effect-). Thus for an ideal gas even this excluded volume cannot be taken for granted. Regardless, if the interaction between two beads on *adjacent* sites is neither repulsive or attractive this will favour mixing, while if the interaction is attractive, this will favour demixing.

Flory-Huggins theory defines three types of interactions. Solute-solute interactions ($\epsilon_{A,A}$), solute-solvent interactions ($\epsilon_{A,0}$), and solvent-solvent interactions ($\epsilon_{0,0}$). Flory-Huggins is a mean-field model, which is to say interactions are determined as the average strength, given composition. Given Flory-Huggins theory is, fundamentally an analytical description of a lattice model, we must also define a coordination number for our lattice (z), which defines to

the number of neighbors associated with each position. In a cubic 3D lattice, a coordination number of 8 or 14 is common.

Much like we first determined the entropy of mixing for a single monomer, the energy of mixing for a single monomer of type A is given by

$$U_A = \phi_A \epsilon_{(A,A)} + \phi_0 \epsilon_{(A,0)} \tag{13.11}$$

This ignores the impact of chain connectivity, and simply determines bead-bead interactions based on the volume-fraction weighted interaction strength of the possible interaction partners (in this case, solute-solute and solute-solvent).

The site specific average energy is then defined as $\frac{zU_A}{2}$, where z is the coordination number (i.e. there are z possible partners for each site) and the 2 is a normalization factor to avoid counting all interactions twice. Recall that $n\phi_A$ describes the number of lattice sites occupied by our polymer, such that the total energy associated with the polymer A is given by

$$U_A^{\text{total}} = \frac{zU_A}{2} \times n\phi_A \tag{13.12}$$

This is simply saying that the total energy of polymer A in this system is given by the energy associated with a single lattice-unit of that polymer, multiplied by the number of lattice-units of that polymer in the system (i.e. U_A^{total} is an extrinsic property).

We treat the solvent in the same manner, and arrive at the following expression for the total energy of the system

$$U = \frac{zn}{2}\Big(U_A\phi_A + U_0\phi_0\Big) \tag{13.13}$$

Because we are describing a binary system we can write ϕ_A as ϕ and ϕ_0 as $(1-\phi)$. This allows us to re-write equation 13.13 as

$$\begin{aligned} U &= \frac{zn}{2}\left\{\Big[\epsilon_{(A,A)}\phi + \epsilon_{(A,0)}(1-\phi)\Big]\phi + \Big[\epsilon_{(A,0)}\phi + \epsilon_{(0,0)}(1-\phi)\Big](1-\phi)\right\} \\ &= \frac{zn}{2}\Big[\epsilon_{(A,A)}\phi^2 + 2\epsilon_{(A,0)}\phi(1-\phi) + \epsilon_{(0,0)}(1-\phi)^2\Big] \end{aligned} \tag{13.14}$$

The expression above gives the energy associated with a fully mixed solution. In the fully demixed state, this becomes

$$U_{\text{demixed}} = \frac{zn}{2}\Big[\epsilon_{(0,0)}(1-\phi)^2 + \epsilon_{(A,A)}\phi^2\Big] \tag{13.15}$$

The change in energy upon mixing is given by

$$\begin{aligned} \Delta U &= \frac{zn}{2}\Big[\epsilon_{(A,A)}\phi^2 + 2\epsilon_{(A,0)}\phi(1-\phi) + \epsilon_{(0,0)}(1-\phi)^2\Big] - \frac{zn}{2}\Big[\epsilon_{(0,0)}(1-\phi)^2 + \epsilon_{(A,A)}\phi^2\Big] \\ &= \frac{zn}{2}\phi(1-\phi)(2\epsilon_{(A,0)} - \epsilon_{(A,A)} - \epsilon_{(0,0)}) \end{aligned} \tag{13.16}$$

Finally, to express this as an intrinsic property we must normalize by the number of sites on the lattice (n).

$$\Delta\bar{U} = \frac{z}{2}\phi(1-\phi)(2\epsilon_{(A,0)} - \epsilon_{(A,A)} - \epsilon_{(0,0)}) \tag{13.17}$$

From this definition we define the Flory χ parameter as

$$\chi = \frac{z}{2}\frac{2\epsilon_{(A,0)} - \epsilon_{(A,A)} - \epsilon_{(0,0)}}{kT} \tag{13.18}$$

Which means we can re-write the energy of mixing per lattice site as

$$\Delta\bar{U} = kT\phi(1-\phi)\chi \tag{13.19}$$

We previously defined the Helmholtz free energy of mixing as $\Delta\bar{F}_M = \Delta\bar{U}_M - T\Delta\bar{S}_M$. We have now defined $\Delta\bar{U}_M$ and $\Delta\bar{S}_M$, meaning we can write our full free energy of mixing expression as

$$\Delta\bar{F}_M = kT\left[\frac{\phi}{N_A}\ln(\phi) + \frac{(1-\phi)}{N_B}\ln(1-\phi) + \chi\phi(1-\phi)\right] \tag{13.20}$$

Or often more conveniently

$$\frac{\Delta\bar{F}_M}{kT} = \frac{\phi}{N_A}\ln(\phi) + \frac{(1-\phi)}{N_B}\ln(1-\phi) + \chi\phi(1-\phi) \tag{13.21}$$

As a quick aside, in the context of solution thermodynamics (based on lattice frameworks) the Gibbs free energy of mixing (ΔG_M) and the Helmholtz free energy of mixing (ΔF_M) are frequently used interchangeably. There is good reason for this. The Helmholtz free energy of mixing provides a thermodynamic description of a system where temperature, number of particles, and volume are fixed. Similarly, the Gibbs free energy of mixing originates from a system where temperature, number of particles, and pressure are held fixed. More generally, we can write the change in enthalpy (ΔH) associated with some process as

$$\Delta H = \Delta U + \Delta pV \tag{13.22}$$

Where ΔU is the internal energy of the process, ΔpV represents a change in volume, pressure, or both. For a lattice at constant temperature, both volume (the number of lattice sites) and pressure (the 'compressibility' of the solvent and solute across the lattice) are held fixed - all sites are occupied by either solvent or solute, and the number of sites is fixed. As a result, the pressure and volume are held fixed, such that for a lattice $\Delta p = 0$ and $\Delta V = 0$. As a result, we can write

$$\Delta H = \Delta U \tag{13.23}$$

Meaning that

$$\Delta\bar{F}_M = \Delta\bar{U}_M - T\Delta\bar{S}_M$$
$$\Delta\bar{F}_M = \Delta\bar{H}_M - T\Delta\bar{S}_M$$
$$\Delta\bar{F}_M = \Delta\bar{G}_M \tag{13.24}$$

Such that in our lattice formalism the Gibbs free energy of mixing ($\Delta\bar{G}_M$) and the Helmholtz free energy of mixing ($\Delta\bar{F}_M$) are formally equivalent to one another.

13.2.4 Polymer Concentration Limits

Flory-Huggins provides a remarkably simple yet incredibly powerful formalism for describing the thermodynamics of mixing. However, it has weaknesses which make it ill suited for fitting certain types of mixing processes. Notably, for polymer solutions where large conformational fluctuations are important, unexpected solution behaviours may arise which Flory-Huggins is unable to capture. To address this, Muthukumar developed a general purpose extension of Flory-Huggins theory that directly takes these fluctuations into account, as well as considering a three-body correction term to address the excluded volume impact of multiple chains [409, 410].

A key idea that becomes critical for Muthukumar theory is that of the **overlap concentration** or overlap volume fraction[27] (c^* or ϕ^*) [504]. Let the radius of gyration (R_G) be the mean radius of gyration and the end-to-end distance (R_{EE}) be the mean end-to-end distance of the polymer of interest. Although R_G and R_{EE} are used interchangeably for describing

[27]The overlap concentration and overlap volume fraction are used interchangeably, as they are directly related by the conversion constant ρ_0

chain statistics and both are relevant for describing solvent-mediated conformational properties of flexible chains, R_G quantifies the internal density of chain units and is a formal order parameter for describing coil-to-globule transitions. Similarly, R_{EE} is the preferred order parameter for quantifying the impact of conformational fluctuations on the pervaded volume of a single chain, which is proportional to R_{EE}^3. In addition R_{EE} is a more appropriate order parameter to quantify the volume occupied by a highly charged polymer due to the long-range electrostatic interactions. The pervaded volume describes the envelope of space occupied by the chain as it thrashes around in solution. This volume is considerably larger than the volume occupied by a single repeating unit. The volume fraction of a single chain inside its pervaded volume is the overlap volume fraction, given by ϕ_A^*. There is a similar definition for the overlap concentration denoted as c_A^*. Assuming a spherical pervaded volume, we obtain the following expression to describe the overlap volume fraction of a polymer of r monomers where the volume of each monomer is v_m.

$$\phi^* = \frac{3rv_m}{4\pi R_{EE}^3} \tag{13.25}$$

The overlap concentration has an important macroscopic meaning; it determines the concentration at which inter-molecular interactions become equally likely as intramolecular interactions (i.e. the concentration that individual polymers begin to 'overlap' with one another).

Polymer solutions can exist in one of three distinct concentration regimes; the dilute, semidilute, and concentrated regimes (see fig. 13.3, which is a reproduction of fig. 12.9).

In the **dilute regime**, the probability of polymers interacting with one another is negligible, and for all intents and purposes they behave as if they are alone in solution. Most measured

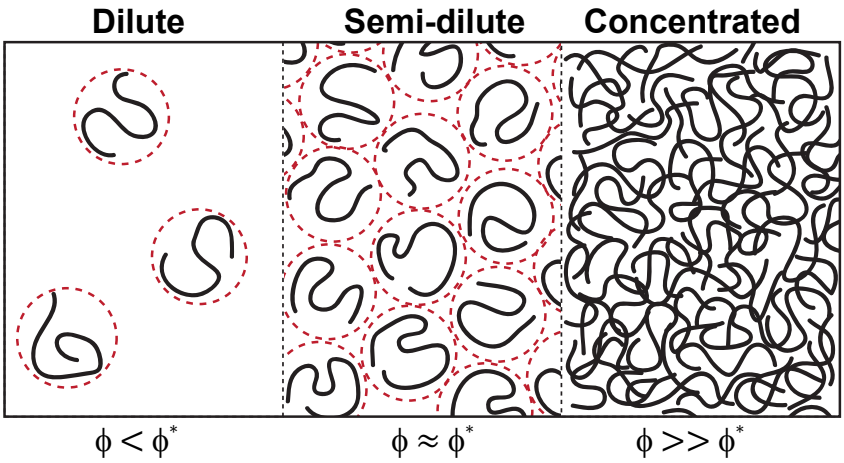

Figure 13.3: Graphical representation of the dilute, semidilute and concentrated regimes. Here, ϕ represents the polymer concentration in the solution and ϕ^* represents the overlap volume fraction. LAF-1 droplets are consistent with a semidilute solution.

thermodynamic properties of polymer solutions in the dilute regime are similar if not identical to those of the pure solvent.

The **semidilute regime** is defined as the concentration equal to and just above the overlap concentration. Here, chains are making frequent intermolecular contacts, but the polymer density is low enough that the polymer remains largely well solvated by solvent. The majority of the volume in the semidilute regime is occupied by solvent. For homopolymers, the interactions and fluctuations will be largely equivalent over the chain, leading to a relatively uniform mesh-size (distance between chains). For heteropolymers, preferential interactions (attractive or repulsive) between distinct positions along the chain are expected to lead to a distinct topological organization within these semidilute solutions. Understanding how the attractive and repulsive nature of chain-chain and chain-solvent interactions dictate the phase behaviour of heteropolymers remains an open but critical question. In the semidilute regime the extent of the conformational fluctuations matter. The average distance between

two points on a chain in the semidilute regime is on the same order of magnitude as the dimensions of the chain itself (this is a necessary condition that originates from our definition of the overlap concentration, and we will return to in subsection 13.2.5). As a result, these fluctuations set a characteristic length scale, such that for polymers that undergo large fluctuations we expect large fluctuations in density in the semidilute regime.

In the **concentrated regime**, the concentration of polymer is so high that polymer-polymer interactions become the major determinants of chain dynamics and conformational behaviour on the monomeric level. Again, while we might expect a concentrated homopolymer solution to form an isotropic environment, preferential interactions in heteropolymers are expected to give rise to local mesoscopic organization in the concentrated regime.

How does the intrinsic conformational behaviour of a polymer influence the overlap concentration? Let us consider a hypothetical homopolymer that shows conventional scaling behaviour according to

$$R_G \propto r^{\nu} \tag{13.26}$$

Where, as before, r is the degree of polymerization[28] and ν is the polymer scaling exponent (as discussed extensively in earlier chapters). ν, like χ, is governed by the balance of chain-chain and chain-solvent interactions. Three characteristic scaling regimes exist for ν; for a poor solvent $\nu \approx 1/3$, for a Θ solvent $\nu \approx 1/2$ and for a good solvent $\nu \approx 3/5$.

How is ν related to the overlap concentration? We can describe the relationship between these three classes of polymers, chain length, and the theoretical overlap concentration. As

[28]We previously denoted r as N, as is more common in the protein biophysics literature

can be seen in fig. 13.4, the more expanded the chain, the lower the overlap concentration. We can also take this to its extreme and imagine the overlap concentration for a fixed rod ($\nu = 1.0$). This has a physically intuitive (and obvious) explanation - as the polymer unit becomes more expanded the overlap concentration becomes lower because the polymer is occupying more space in the solution, where 'space' is quantified in terms of the pervaded volume.

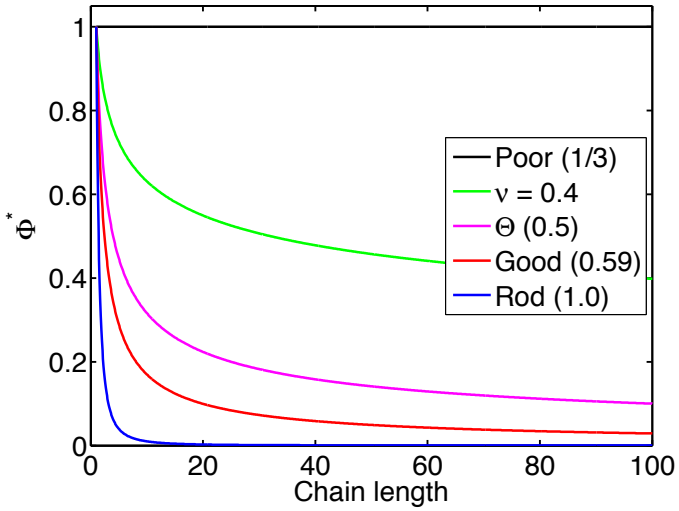

Figure 13.4: Relationship between chain length (r), intrinsic scaling exponent (ν), and overlap concentration (ϕ^*) for a non-interacting chain, based on equation 13.27. As the dimensions of the individual chain become larger the overlap concentration decreases. The value in parenthesis in the legend reflects the ν used to generate the curve. Note that this reflects the scaling relationship, although absolute numbers will depend on system specific prefactors

How does the overlap concentration relate to the phase boundary associated with demixing? Based on the usFCS and 3D confocal microscopy measurements of LAF-1 droplet concentrations we know the dense phase of the LAF-1 droplet is equivalent to a semidilute solution. The concentrations are simply too low to represent the concentrated solution, and the dense phase of a two-phase system cannot be in the dilute phase as this by definition requires an

absence of polymer-polymer interactions. Consequently, we expect the fluctuations inside the droplet to be on a similar order of magnitude to the dimensions of the LAF-1 molecules themselves. Secondly, the low concentration arm of the binodal curve is extremely low, with phase separation occurring at a volume fraction of ≈ 0.0002 (0.02%). This raises a conundrum that Flory-Huggins theory is unable to explain; how can a polymer undergo phase separation at such low concentrations (implying strong intermolecular interactions), yet give rise to such a dilute dense phase? We can explain this result using Muthukumar's theory of polymer solutions, which despite its complexity has a simple and somewhat intuitive interpretation, as will be discussed later.

Before we continue, it is important to address an obvious issue when we consider intrinsic chain behaviour, the overlap concentration and phase separation. Figure 13.4 shows that as a polymer becomes more extended, the overlap concentration decreases. A naive interpretation of this is that the more expanded the polymer, the lower the concentration phase separation will occur at. For homopolymers, chain expansion *necessarily* reflects a reduction in the χ parameter. For homopolymers, inter and intramolecular interactions are equivalent, such that for polymers with a very low overlap concentration (negative χ) phase separation is unobtainable because there is no driving force for intermolecular interactions. Similarly, for homopolymers for which water is a poor solvent (strong positive χ), the high overlap concentration is entirely superseded by the system's driving force to minimize the exposed surface area of the polymer, leading to rapid demixing. We strongly caution that for homopolymers the overlap concentration is simultaneously a measure of the concentration at which chains begin to overlap, and the strength of solute-solute interactions. Given that these two factors promote and inhibit, respectively, intermolecular interactions, caution should be exercised when interpreting how the overlap concentration of a given polymer may (or may not) provide insight into its ability to phase separate.

13.2.5 The Effective Scaling Exponent and the Correlation Length

Like homopolymers, for heterpolymeric sequences we can quantify the average chain behaviour according to some *effective* scaling exponent. We will write this scaling exponent as ν for simplicity, but as discussed extensively in chapter 7 the notion of a complex heteropolymer such as an IDP behaving in a manner were the scaling exponent provides local and global insight can be misleading (and simply incorrect). We can describe the overlap volume fraction is a function that depends on ν

$$\phi^* \sim r^{1-3\nu} \tag{13.27}$$

As ν becomes larger, there is an increasingly sharp decrease in ϕ^* as a function of degree of polymerization (r). IDPs are biological heteropolymers, and for the purpose of our discussion can be separated into two distinct classes depending on the intrinsic conformational behaviours (as described by ν). One class of IDPs exits where ν lies between 1/2 and 3/5 (between a polymer at the Θ solvent and good solvent limits, i.e. a generic good solvent). A second class exists where ν lies between 1/2 and the poor solvent limit of 1/3 (i.e. a generic poor solvent).

For IDPs in the first class to phase separate requires that the chain consist of strongly segregating regions that are unable to satisfy one another via intramolecular interactions due to the corresponding entropic penalty, but in the context of a phase separated droplet engage in extensive inter-molecular interactions. This type of phase separation is simply not obtainable via homopolymers, but we hypothesize is a mechanism through which dilute droplets can be realized.

For the second class chain-chain interactions are on average stronger than chain solvent interactions, allowing phase separation via mechanisms more consistent with a homopolymer. It is worth pointing out that due to their heteropolymeric nature, the relationship between ν and phase separation need not be predictive. A prime example of this is folded proteins, which typically have a ν of around 1/3 but most of which are soluble up to relatively high concentrations [144]. A key theme that will re-emerge later is the idea that heteropolymers allow for a decoupling between intra and inter molecular interactions in a manner than homopolymers do not.

Beyond the overlap volume fraction ($\phi > \phi^*$) in the semidilute regime there is an increasing probability that chains will become part of a network with other chain molecules. At a given concentration (or volume fraction) of polymer in solution, we have to account for the typical distance between a pair of contact points taken from chains (note that this could be two points on the same chain, or two points taken from different chains). This characteristic distance is known as the mesh size [3,133]. In semidilute solutions for length scales below the mesh-size the polymers solution behaviour is dictated by its local environment, including attractive and repulsive interactions and excluded volume effects. For length-scales above the mesh size these types of interactions are screened by solute and solvent, giving rise to a chain that over these greater length scales behaves like a Flory random coil ($\nu = 1/2$). Consequentially in the semidilute regime the correlation length (ξ), also known as the screening length, is equivalent to the mesh size.

The correlation length (as defined by de Gennes) is set to be [133]:

$$\xi \sim R_G \left(\frac{\phi}{\phi^*} \right)^x \tag{13.28}$$

Where x is defined as

$$x = \frac{\nu}{1 - 3\nu} \tag{13.29}$$

Recall that we can write

$$R_G = br^{\nu} \tag{13.30}$$

Where b is a fixed prefactor which for unfolded polypeptides was determined to be $\sim$1.9 [297]. It should be noted that frequently it is assumed that this prefactor is an approximately fixed value, but our work in chapter 7 suggests this prefactor varies significantly as a combined function of chain length, apparent solvent quality, amino acid sequence, and preferential long-range and short range interactions. Moreover, in the original derivation of the $R_G = br^{\nu}$ relationship, this prefactor is expected to be approximately solution independent only when $\nu > 0.5$. Taken together, equation 13.28 can be re-cast in an expanded form

$$\xi \sim br^{\nu} \left(\frac{\phi}{\phi^*} \right)^{\left[\frac{\nu}{1 - 3\nu} \right]} \tag{13.31}$$

From this, it should be clear that the correlation length (ξ) depends on the *relative* polymer concentration (where that relativity references the overlap volume fraction ϕ^*), the effective scaling exponent (ν), the degree of polymerization (r) and the prefactor term (b). For solutions in the semidilute ($\phi \approx \phi^*$) or concentrated ($\phi >> \phi^*$) regimes ξ describes the lengthscale over which density fluctuations are felt. Figure 13.5 examines how the correlation changes as a function of chain-length and solution density, illustrating that how the screening length decreases as polymer concentration increases.

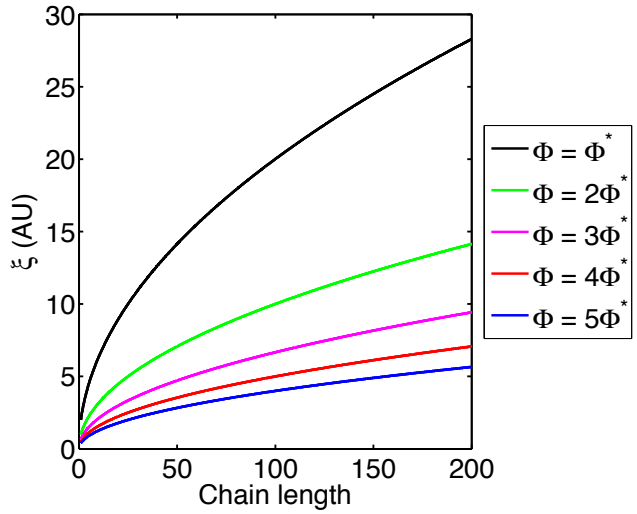

Figure 13.5: Relationship between chain length (r), solution concentration (ϕ), and the correlation-length ξ for a non-interacting chain, based on equation 13.31. As solution concentration increases the correlation length decreases. Note that again we are only interested in the scaling behaviour here, not the absolute values (which are in arbitrary units).

In the semidilute regime $\phi \approx \phi^*$, such that

$$\begin{aligned} \xi &\sim R_G \left(\frac{\phi}{\phi^*} \right)^x \\ &\sim R_G \Big(1 \Big)^x \\ &\sim R_G \end{aligned} \tag{13.32}$$

This provides an important equivalence that we will return to later - the ensemble average global dimensions of a monomer in the semidilute regime should be approximately the same order of magnitude as the correlation length. Note that this is true far from the critical

point - as we approach the critical point this relationship breaks down and can no longer be expected to hold true, given that fluctuations grow to become system spanning as the critical point is approached.

A final and important concept to introduce at this juncture is that of the blob. Unhelpfully, there are two types of blobs of interest to us, the *thermal blob* and the *concentration blob*.

The **thermal blob** is primarily of interest in the limit of chains in the dilute concentration regime, but more generally refers to a length scale that is intrinsic to an individual polymer. It describes the number of monomers over which the chain's conformational perferences are approximately the same as a Gaussian chain, with chain-chain and chain-solvent interaction strengths being on the same order of magnitude as thermal fluctuations (i.e. $\sim kT$). For polypeptides there is some weak sequence dependence on the thermal blob, but by-and-large this value is approximately 5-6 residues [126]. If we define this length-scale as g_b (in units of number of amino acids), then the radius of gyration associated with a thermal blob can be written as

$$R_G^b \approx R_0^b g_b^{0.5} \tag{13.33}$$

R_0^b represents a prefactor with the same physical meaning as the prefactor b we discussed previously, although the numerical value associated with R_0^b will differ from the equivalent prefactor used for a full polymer. In previous work we found that R_G^b is approximately ~ 6.0 Å.

The **concentration blob** is primarily of interest in the limit of chains in the semidilute or concentrated regimes. The concentration blob is the length-scale associated with selecting a thermal blob on one chain and then picking a nearby thermal blob (irrespective of if that second blob comes from the same chain or a different chain). The average distance

between neighbouring thermal blobs reflects the size of the concentration blob, which is entirely equivalent to the correlation length/screening length (ξ) in the semidilute regime, as described previously. A useful idea for thinking about the concentration blob is g_ξ, which describes the dimensions of the concentration blob in terms of the number of residues that contribute to the concentration blob. We can map this back into real spatial dimensions by dividing by the number of residues in a thermal blob (which gives the number of thermal blobs in a concentration blob) and then multiplying by the dimensions of a thermal blob.

As a result

$$\begin{aligned}\xi &\sim \frac{g_\xi}{g_b} R_G^b \\ &\sim \frac{g_\xi}{5} \times 0.6 \\ &\sim 0.12 g_\xi \end{aligned} \tag{13.34}$$

Here 0.12 is in units of nm per residue, and reflects the contribution each residue makes to the screening length. In Muthukumar's original formulation g_ξ (there written as ξ, which we have reserved in this work to describe the correlation length in units of nanometers) is given in units of Kuhn lengths, and the degree of polymerization (r in our notation) is given in number of Kuhn lengths (n in the original formalism). As a result, we are effectively re-setting the length-scale for our interpretation of the theory into the units of residues, rather than the units of Kuhn lengths. This is convenient, but also formally correct (the best kind of correct). The Kuhn length is typically described as half of the persistence length, where the persistence length is the length-scale over which the polymer behaves as a rigid rod. For

proteins the persistence length is necessarily equal to a single residue, given the flexibility of the amide bonds. As a result, if we take the persistence length to be ~ 0.3 nm (the generally used contour length associated with a single amino acid) then the Kuhn length emerges as 0.15, almost identical the value we calculate above.

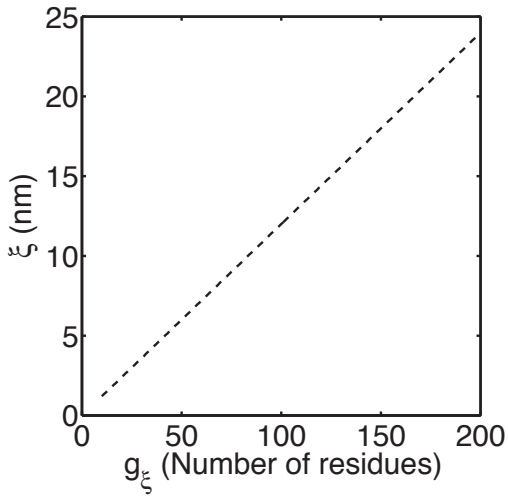

Figure 13.6: Conversion between g_ξ and ξ

The clear implication from this is that the correlation length (ξ) depends on the protein concentration, which in this case is given by the number of residues that contribute to the concentration blob (g_ξ). The mapping between ξ and g_ξ is shown in fig 13.6.

13.2.6 Muthukumar's Theory of Polymer Mixing

Muthukumar's theory of polymer mixing combines the standard Flory-Huggins expression for the free energy of mixing with a three-body correction term and a density fluctuations term [409, 410]. As derived previously, the Flory-Huggins free energy of mixing originates from a combination of the enthalpy/energy of mixing with the entropy of mixing, giving rise to the following expression;

$$\frac{\Delta G_{FH}}{k_B T} = \frac{\phi}{r} \ln \phi + (1-\phi) \ln(1-\phi) + \chi \phi (1-\phi) \tag{13.35}$$

de Gennes extended this description by adding a three-body correction term that relies on w [133];

$$\frac{\Delta G_{FH\&G}}{k_B T} = \frac{\phi}{r} \ln \phi + (1-\phi) \ln(1-\phi) + \chi \phi (1-\phi) + \left(w - \frac{1}{6} \right) \phi^3 \tag{13.36}$$

While χ is proportional to the second virial coefficient (B_2) and reflects a correction to ideal-gas behaviour associated with binary interactions, w is proportional to the third virial coefficient (B_3) and reflects a correction to account for three body interactions. Can we develop a more physical intuition as to what these somewhat opaque three-body interactions are? Imagine three spheres, all of which are uniformly attracted to one another via a binary (two-body) interaction potential, attempting to interact with one another in a ternary complex. In the absence of a three-body correction term these spheres can only 'feel' one partner, but not the other, leading to a physical overlap of the spheres. The three-body correction allows for this ternary interaction to occur in a physically realistic manner, and

effectively becomes an excluded volume correction term for the three-way-intersection of solute components. In effect, the three-body interaction dilutes the attractive power of the two-body interaction in a bulk-concentration dependent manner.

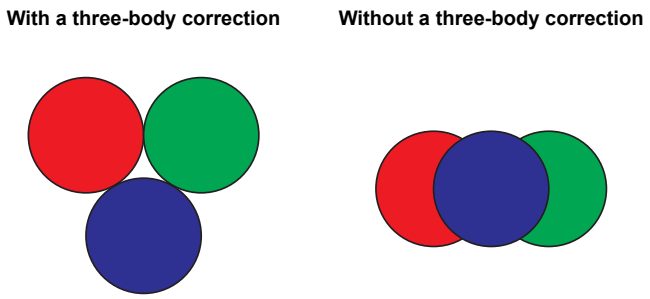

Figure 13.7: Graphical description of what the three-body correction term (w) means

Muthukumar's definition of the free energy of mixing builds on the work of Flory and de Gennes, and further uses the three body interaction term to define how conformational fluctuations influence the free energy of mixing. In addition to w, the Muthukumar free energy of mixing includes two additional parameters, α and g_ξ, and is defined as

$$\frac{\Delta G_M}{k_B T} = \frac{\phi}{r}\ln\phi + (1-\phi)\ln(1-\phi) + \chi\phi(1-\phi) + \left(w - \frac{1}{6}\right)\phi^3 + \frac{1}{24\pi g_\xi^3} - \frac{9}{16\pi}\frac{(1/2 - \chi + w\phi)\phi}{\alpha^2 g_\xi} \tag{13.37}$$

α defines the swelling ratio, a parameter that provides information on how the chain's conformational behavior changes upon entry in to the polymer-rich regime from the polymer-poor regime. g_ξ defines the concentration-dependent correlation length as discussed in the preceding sections. Before further unpacking Muthukumar's theory of polymer solutions, we shall overview how one converts a free energy of mixing curve into a phase diagram.

13.2.7 One-Phase vs. Two-Phase Stability

Despite the varying complexities associated with these different expressions for the free energy of mixing, they are all asking a fairly (conceptually) simple question: what is the free energy associated with the fully mixed (one-phase) state given some specific volume fraction of polymer? For a combination of fixed parameters (χ, r *etc.*) one can construct a free energy of mixing curve for all values of ϕ as shown in figure 13.8 below;

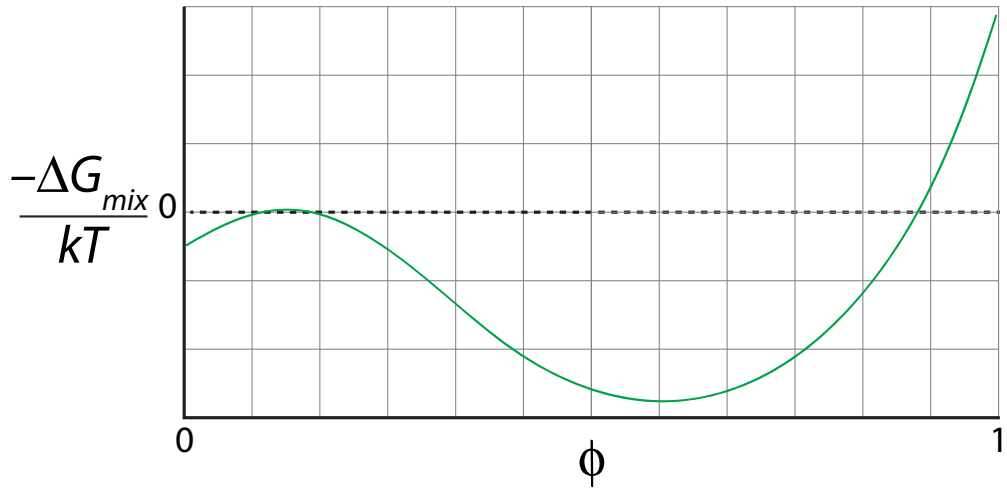

Figure 13.8: Example free energy of mixing curve. This curve was generated by evaluating the free-energy of mixing function monotonically and equally spaced values of ϕ (e.g. $\phi = 0.001,\ 0.002,\ 0.003,\ ...,\ 0.999$). Note that such a curve is generated by fixing the other parameters (e.g. $\chi, N, w, g_\xi, \alpha$) and *only* varying ϕ.

From such a free energy of mixing curve we can determine (for the set of defined parameters used to generate this particular curve) the volume regions where the mixed system is *stable* (the one-phase regime is favoured), *metastable* (the two-phase is favoured but a one-phase system is stable to standard fluctuations associated with the system), and *unstable* (the two-phase regime is favored and is realized immediately).

To extract the volume fraction values from the free energy of mixing curve that correspond to the location of the metastable (binodal) and stable (spinodal) points on a phase diagram is somewhat non-trivial, and requires additional explanation. In the section that follows we outline, formally, how one determines the binodal and spinodal points associated with a given free energy of mixing curve. These ideas can then be used to construct a complete phase diagram by determining how the binodal and spinoda points change as a function of χ from multiple free energy of mixing curves.

For each point on the free energy curve, we wish to determine if the mixed (one-phase) regime is more stable or less stable than *any* phase separated (two-phase) regime. At this point we will not concern ourselves with the specific composition of the two-phase system. Any mixed (one-phase) volume fraction on the free energy of mixing curve can be re-written as a demixed (two-phase) system where one of the two-phases has a polymer volume fraction above the fully mixed polymer volume fraction and the other a polymer volume fraction below the original mixed composition. This is possible because we are not constraining the total fraction of the polymer that goes into each of those two-phases, only that the volume-weighted polymer fraction of those two phases must add up to the total fraction of polymer in the mixed system.

To help illustrate this idea, consider a system with ANY polymer fraction. This system can be re-configured into a two-phase system where one-phase is pure polymer ($\phi_2 = 1.0$) and one-phase is pure solvent ($\phi = 0.0$). The volume of the system occupied by each of these two-phases will depend on the overall polymer fraction; for an example of this and a number of other scenarios see figure 13.9, which demonstrates different ways a mixed system can be re-arranged into different two-phase systems.

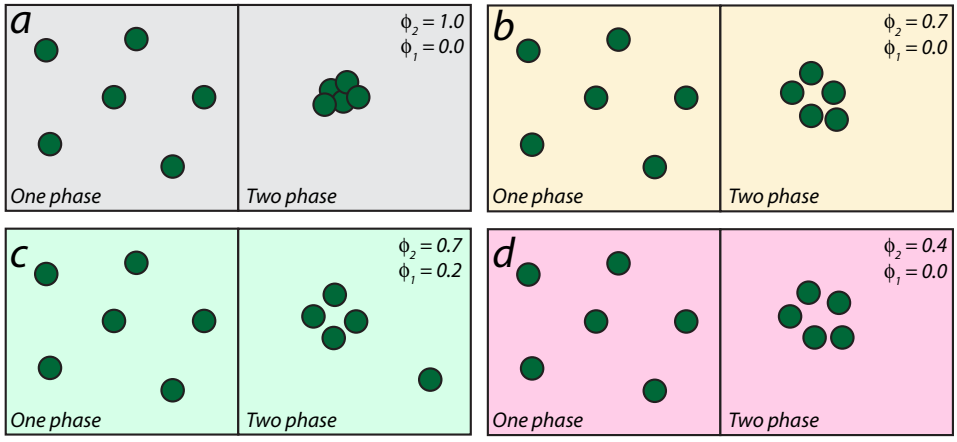

Figure 13.9: Four possible examples of a mixed system undergoing phase separation into four distinct two-phase regimes. We use the convention here that ϕ_1 refers to the polymer volume fraction in the dilute phase while ϕ_2 refers to the polymer volume fraction in the dense phase.

This conceptually illustrates how - within some specific constraints outlined below - the system *can* be reconfigured into an infinite number of different two-phase regimes. While this describes the fact that the system *can* demix it does not necessarily mean the system *will* demix; this depends on the underlying energetic of the demixed system vs. the fully mixed system.

To add some quantitative rigor, the free energy of the one-phase regime at volume fraction ϕ_M is written as

$$\Delta G_{mix}^{1 \text{ phase}}(\phi_M) = \Delta G_{mix}(\phi_M) \tag{13.38}$$

While the free energy of the two-phase regime that comes from that same bulk volume fraction can be written as

$$\Delta G_{mix}^{2 \text{ phase}}(\phi_M) = \alpha \Delta G_{mix}(\phi_1) + \beta \Delta G_{mix}(\phi_2) \tag{13.39}$$

Where the following constraints apply

$$\phi_1 < \phi_M \;\; \& \;\; \phi_2 > \phi_M \tag{13.40}$$

i.e., the two-phases in a two-phase system have a polymer fraction greater than and less than the mixed regime.

Additionally,

$$\alpha\phi_1 + \beta\phi_2 = \phi_M \tag{13.41}$$

This simply describes the fact that we are not creating or destroying polymer, but partitioning it into specific regions with specific volume fractions.

Finally,

$$\alpha + \beta = 1.0 \tag{13.42}$$

α and β represent the relative fraction of the system which is in each of the two-phases - i.e. this is simply a *lever rule*. Note that - importantly - we do not put any constraint on the value of α or β.

Having established how our one-phase system *can* re-configure itself into a two-phase system, we next must ask if the two-phase regime is more energetically favorable than the mixed regime. Graphically, we can determine the free energy of a two-phase system by drawing a tie line between the two volume fractions associated with our two two-phase concentrations (e.g. the tie-line connects the volume fraction in the dilute phase (ϕ_1) with the volume fraction in the dense phase (ϕ_2)).

The tie line describes the correctly weighted free energy of the two-phase regime at any starting volume fraction for the mixed regime. Figure 13.10 illustrates this idea with an example. The one-phase regime has a polymer volume fraction of ϕ_M, while in our hypothetical two-phase regime the dilute phase has a volume fraction of ϕ_1 (1) while the concentrated phase has a volume fraction of ϕ_2 (2).

The free energy associated with this system in the two-phase regime $\left(\Delta G_{mix}^{2 \text{ phase}}\right)$is higher than the free energy associated with the mixed regime $\left(\Delta G_{mix}^{1 \text{ phase}}\right)$. As a result, we know that the one-phase system of this composition system will not decompose into *this specific* two-phase configuration, because the one-phase configuration is more energetically favourable.

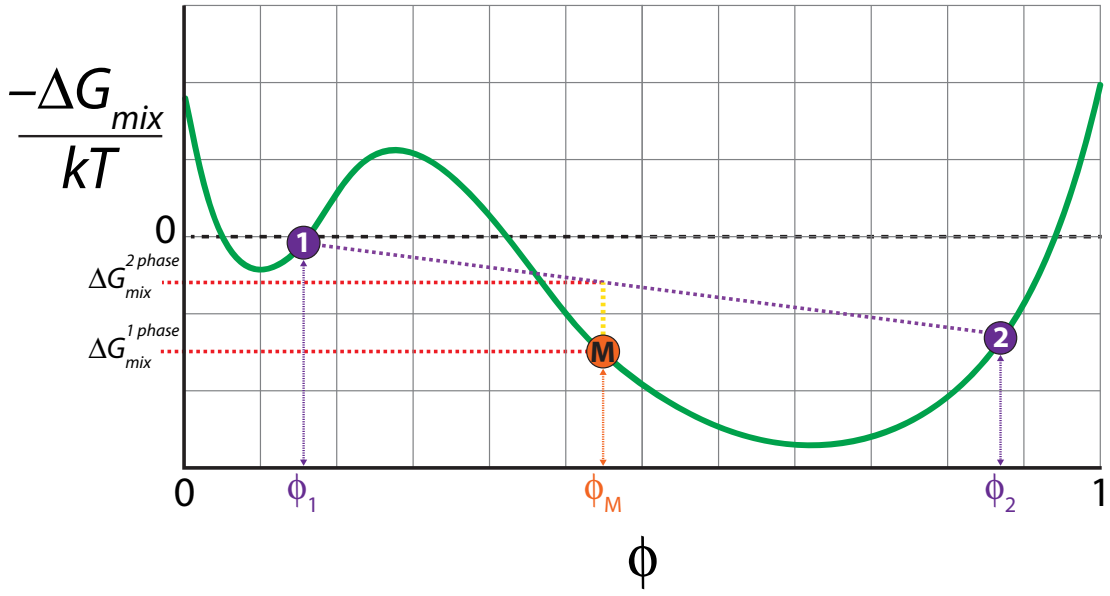

Figure 13.10: Demonstration of how we might compare the relative stability of a one-phase vs. a *specific* (and arbitrary) two-phase configuration, where the dense phase has a polymer fraction of ϕ_2 and and the dilute phase has a polymer fraction of ϕ_1. Note that to determine the free energy of mixing associated with the two-phase system for ANY initial polymer volume fraction between ϕ_1 and ϕ_2 a vertical line is drawn between the specific initial mixed volume fraction (which in this case is ϕ_M) and the tie line. The intersection of this line with the tie-line defines the free energy of the two-phase system.

However, this does not provide any information regarding the *general* stability of the one-phase regime associated with a polymer volume fraction of ϕ_M, but only provides a relative comparison of the one-phase regime and a specific (and entirely arbitrary) two-phase configuration. We could have drawn the tie lines between any two points on the curve (providing $\phi_1 < \phi_M$ and $\phi_2 > \phi_M$). Considering this, we want a *general* way to evaluate the relative stability of all two-phase regimes vs. the fully mixed one-phase regime.

Conveniently, the general solution to this question can be obtained by drawing the *common tangent* lines on the phase diagram. Such a line (the purple dashed line in Figure 13.11) connects two positions on the curve which share a common tangent line, which must by definition (given the general convexity of free energy curves) define the minimum energy line across the surface. Practically, this means that any point on the free energy surface that lies above this line is more favourable in the two-phase regime. This also means that two points connected by the common tangent are the binodal values for the χ value used to generate this specific free energy of mixing curve. Similarly, the points at which the second derivative of the free energy = 0 are the spinodal points for this χ value. We can connect this back to our physical intuition of what a two-phase system is. The concentrations in the dense-phase and dilute phase represent fixed points in some phase space. Upon the addition of some solute of interest, the system will be out of equilibrium with respect to the chemical potential, but will re-equilibrate by changing the volumes of the two phase regions such that their concentrations return back to the fixed points on the free energy surface.

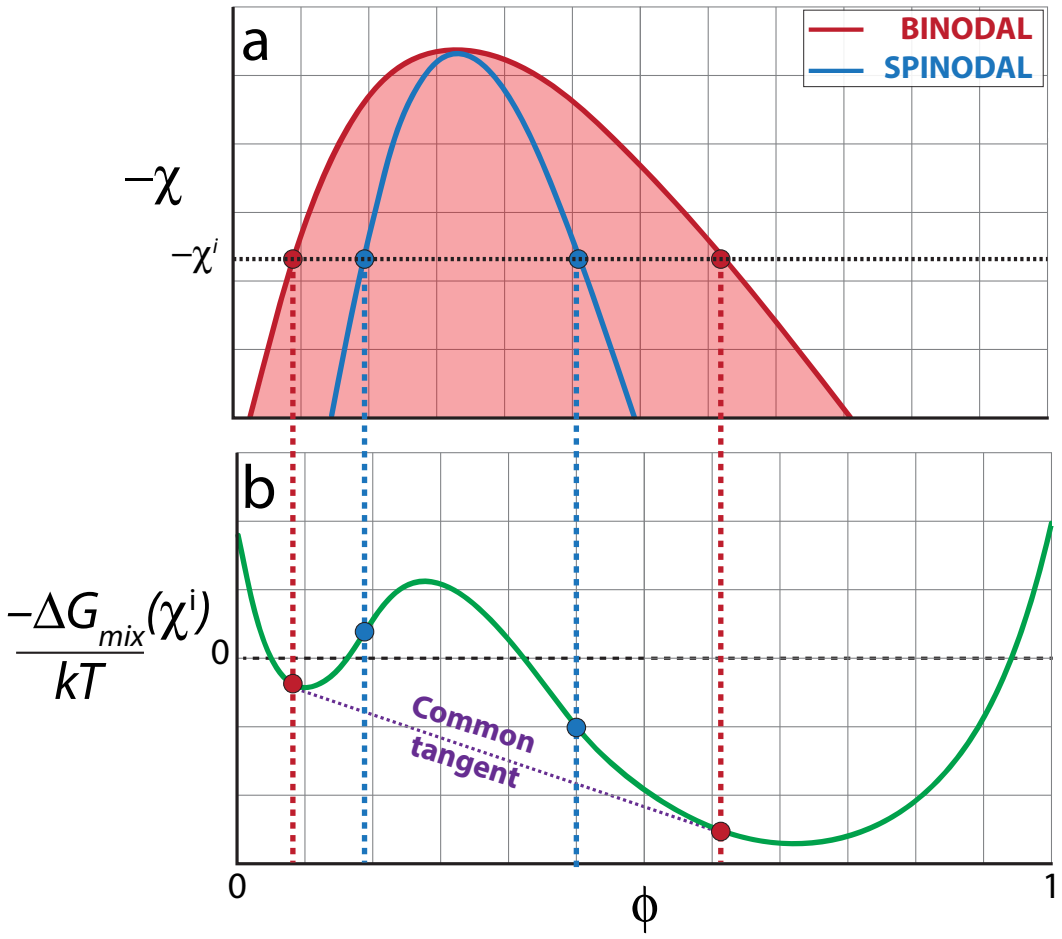

Figure 13.11: Panel b shows the free energy curve generated with some specific χ value (χ^i). As a result, the spinodal and binodal points shown on the phase diagram along the coexistence curve (panel a) are specific for that χ value

Given this, it should now become clear how one uses the free energy expressions to construct a χ vs. ϕ phase diagram. We generate *many* free energy curves by systematically varying χ, and for each curve solve to determine the binodal and spinodal values. These values are then used to construct the χ vs. ϕ phase diagram. The second derivatives are solvable analytically, but the common tangent must be solved numerically by determining points on the curve where the gradients match one another (or a number of specific edge cases,

outlined below). The fact that spinodal values can be determined analytically is a major advantage, as it provides a built in sanity check - every pair of binodal values must have a corresponding pair of spinodal values, which are always within the bounds of the binodal values. This provides an (effectively) free safety check which we use to ensure that our numerical implementation for the common tangent solution provides us with reasonable solutions.

The free energy of mixing surface defined by the Muthukumar free energy of mixing has a number of features that are simply not observed in Flory-Huggins theory. One aspect of this not addressed elsewhere is that while the common tangent is the solution to the binodal points, the solution to the biondal points is not necessarily the common tangent. Put another way - if we can draw a common tangent, that common tangent will define the binodal points, but if we cannot draw a common tangeng this does not *necessarily* mean there are no binodal points. This seems like a tautologous and subtle point, but has important implications from the perspective of a numerical algorithm for identifying the binodal values. Take the example in Figure 13.12, where our free energy curve is uniformly concave with a single maxima. Such a system apparently has no spinodal points (the second derivative of the free energy is never 0), and given there is no common tangent one might expect such a system will not phase separate. In fact, any volume fraction of polymer will lead to a perfectly separated two-phase configuration, where the dilute phase has a $\phi = 0.0$ while the dense phase has $\phi = 1.0$. The purple line in figure 13.12 is not the common tangent, but instead represents the free energy surface minimization with respect to ϕ - i.e., the free-energy tie line that is uniformly equal to or below the free energy curve. This illustrates the fact that in utilizing the free energy of mixing curve, our objective is to identify the tie line that gives rise to an energy-minimized free energy surface. The common tangent will achieve this, but there is nothing fundamentally important about the solution being the common tangent.

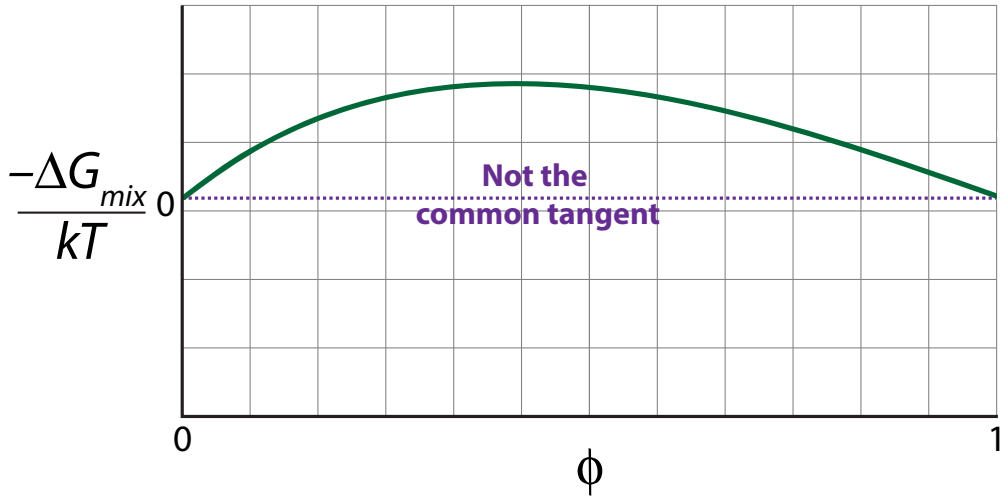

Figure 13.12: Free energy of mixing curve for a system composed of a polymer with an extremely positive χ value (i.e. a polymer in an extremely poor solvent). The two-phase regime is favoured at all polymer volume fractions

With this in mind, determining the binodal points associated with a free energy of mixing curve involves a number of steps, where the correct procedure depends on the underlying structure of the associated free energy of mixing curve. In this work, we have developed a general purpose algorithm which combines discrete information from the first and second derivatives of the free energy of mixing (i.e., Maxwell construction) with additional edge cases to deal with scenarios where a minimal free energy tie line can be constructed that is not a common tangent line. This algorithm can be applied to different theoretical definitions of the free energy of mixing, such as Flory-Huggins or Muthukumar, but is independent of the actual free energy of mixing definition.

The remainder of this chapter is organized as follows:

1. For completeness, we include the full free energy of mixing expressions and their first, second, and third derivatives - these represent the analytical framework upon which our numerical apparatus operate

2. We introduce some of the general and practical challenges associated with generating theoretical phase diagrams to match experimental data

3. Finally we discuss the fitting of a theoretical curve to the full LAF-1 phase diagram, outlining the general steps taken and the origin of the various constants, parameters and assumptions made.

13.3 Free Energy of Mixing Derivatives

The following expressions are included here for completeness.

13.3.1 Flory-Huggins

The full Flory-Huggins free energy of mixing expression is as follows

$$\frac{\Delta G_{FH}}{k_B T} = \frac{\phi}{r}\ln\phi + (1-\phi)\ln(1-\phi) + \chi\phi(1-\phi) \tag{13.43}$$

The first derivative of the Flory-Huggins free energy is

$$\frac{\partial}{\partial\phi}\Big(\frac{\Delta G_{FH}}{k_B T}\Big) = -\ln(1-\phi) + \frac{1}{r}\ln\phi - 1 + \frac{1}{r} + (1-2\phi)\chi \tag{13.44}$$

The second derivative of the Flory-Huggins free energy is

$$\frac{\partial^2}{\partial\phi^2}\Big(\frac{\Delta G_{FH}}{k_B T}\Big) = \frac{1}{r\phi} + \frac{1}{1-\phi} - 2\chi \tag{13.45}$$

The third derivative of the Flory-Huggins free energy is

$$\frac{\partial^3}{\partial\phi^3}\Big(\frac{\Delta G_{FH}}{k_B T}\Big) = \frac{1}{(1-\phi)^2} - \frac{1}{r\phi^2} \tag{13.46}$$

13.3.2 Flory-Huggins with Three Body Correction (w)

The full Flory-Huggins free energy of mixing with the three body correction is as follows

$$\frac{\partial}{\partial\phi}\Big(\frac{\Delta G_{FH\&G}}{k_BT}\Big) = \frac{\phi}{r}\ln(\phi) + (1-\phi)\ln(1-\phi) + \chi(1-\phi)\phi + (w-\frac{1}{6})\phi^3; \tag{13.47}$$

The first derivative of the Flory-Huggins free energy of mixing with the three body correction is

$$\frac{\partial}{\partial\phi}\Big(\frac{\Delta G_{FH\&G}}{k_BT}\Big) = -\ln(1-\phi) + \frac{1}{r}\ln\phi - 1 + \frac{1}{r} + (1-2\phi)\chi - \frac{\phi^2}{2} + 3w\phi^2 \tag{13.48}$$

The second derivative of the Flory-Huggins free energy of mixing with the three

$$\frac{\partial^2}{\partial\phi^2}\Big(\frac{\Delta G_{FH\&G}}{k_BT}\Big) = \frac{1}{r\phi} + \frac{1}{(1-\phi)} - 2\chi + \phi(6w-1) \tag{13.49}$$

The third derivative of the Flory-Huggins free energy of mixing with the three

$$\frac{\partial^3}{\partial\phi^3}\Big(\frac{\Delta G_{FH\&G}}{k_BT}\Big) = \frac{1}{(1-\phi)^2} - \frac{1}{r\phi^2} + 6w - 1 \tag{13.50}$$

13.3.3 Muthukumar Free Energy of Mixing

The full Muthukumar free energy of mixing is defined as follows

$$\frac{\Delta G_M}{k_B T} = \frac{\phi}{r}\ln\phi + (1-\phi)\ln(1-\phi) + \chi\phi(1-\phi) + \left(w - \frac{1}{6}\right)\phi^3 + \frac{1}{24\pi g_\xi^3} - \frac{9}{16\pi}\frac{(1/2 - \chi + w\phi)\phi}{\alpha^2 g_\xi} \tag{13.51}$$

The first derivative of the Muthukumar free energy of mixing is defined as follows

$$\frac{\partial}{\partial\phi}\Big(\frac{\Delta G_M}{k_B T}\Big) = -\ln(1-\phi) + \ln\phi\frac{1}{r} - 1 + \frac{1}{r} - \phi\chi + (1-\phi)\chi + 3\phi^2(w - \frac{1}{6}) - \frac{9}{16\pi}\frac{(1/2 - \chi + \phi w)}{\alpha^2\xi} - \frac{9}{16\pi}\frac{\phi w}{\alpha^2\xi} \tag{13.52}$$

The second derivative of the Muthukumar free energy of mixing is defined as follows

$$\frac{\partial^2}{\partial\phi^2}\Big(\frac{\Delta G_M}{k_B T}\Big) = \frac{1}{r\phi} + \frac{1}{1-\phi} - 2\chi + \phi\Big(6w - 1\Big) - \frac{9w}{8\alpha^2\pi\xi} \tag{13.53}$$

And finally, the third derivative of the Muthukumar free energy of mixing is given as follows

$$\frac{\partial^3}{\partial\phi^3}\Big(\frac{\Delta G_M}{k_B T}\Big) = \frac{1}{(1-\phi)^2} - \frac{1}{r\phi^2} + 6w - 1 \tag{13.54}$$

13.4 Phase Diagrams from Free Energy of Mixing Curves

The general approach for the construction of phase diagrams in the χ vs. ϕ space is outlined below.

1. A free energy of mixing curve is generated across all volume fractions of polymer (0 to 1).

2. The binodal points for that specific free energy of mixing curve are determined - there are two points (low concentration arm and high concentration arm) which correspond to coexistence points associated with the χ used to generate that specific free energy surface. The details associated with the identification of the binodal points are dealt with in subsection 13.4.1.

3. Simultaneously, if applicable the spinodal points are computed analytically as the points where the second derivative of the free energy of mixing are zero

4. This process is repeated with a range of χ values. For every unique χ value, two points on the coexistence line (binodal) are generated and if applicable two points on the spinodal curve are also generated.

5. The smoothness of the phase diagram depends on the number of χ values used.

Figure 13.13 shows this procedure graphically. Note that the free energy curves are offset from one another to improve readability (in reality they would all be close to overlapping).

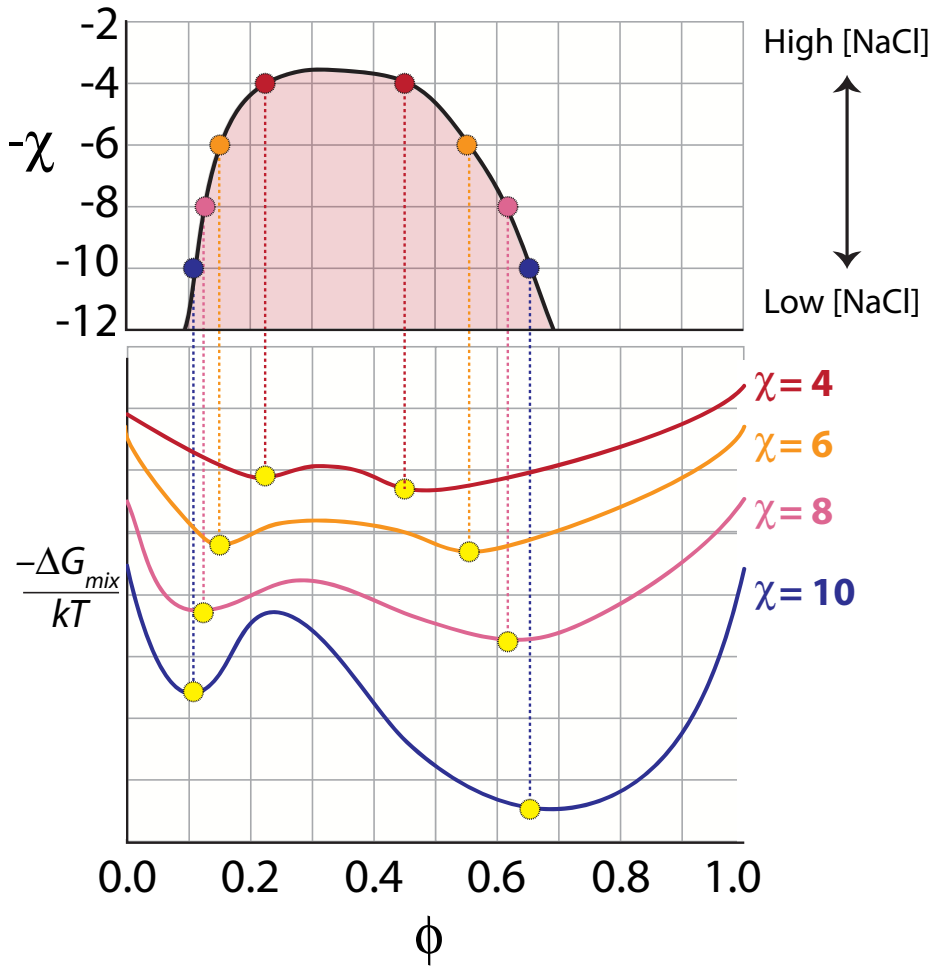

Figure 13.13: Practical schematic showing how multiple free energy of mixing curves correspond to multiple points on the a χ vs ϕ phase diagram. Each curve on the bottom panel corresponds to the free energy of mixing at a different χ value, as shown by the legend. The common tangent points from those curves then define the corresponding points on coexistence curve in the top panel.

13.4.1 Practical Numerical Issues

What are the numerical challenges associated with the constriction of these phase diagrams? One consideration which we should introduce early on is the numerical cost of solving the

common tangent. To determine the common tangent, we determine two positions on the curve where the gradient (first derivative of the free energy) and the x axis intercept are identical (i.e two positions on the curve that are also two positions on the same line). To identify two positions where the gradients match one another we must compare the gradients at two specific volume fractions. We can take advantage of the fact we know the low concentration arm binodal must be below or equal to the low concentration arm spinodal and, similarly, the high concentration binodal must be equal to or above the high concentration arm spinodal (recall the spinodals are the two solutions to equation 13.55[29]).

$$\frac{\partial^2 \Delta G}{\partial \phi^2} = 0 \tag{13.55}$$

This limits our search domain for gradient comparisons, reducing the computational cost. Despite this, we are left with the challenge of comparing discrete points on the curve, where the **volume fraction resolution** influences both the accuracy of the comparison but also the computational cost of the search. The volume fraction resolution refers to the number of discrete points between $\phi = 0.0$ and $\phi = 1.0$, where for each point the first and second derivative is calculated. To illustrate the importance of the volume fraction resolution see figure 13.14.

In this figure, the full free energy of mixing is shown in green, but we have chosen (arbitrarily) to evaluate the derivative with a volume fraction resolution of 12 (i.e. 13 independent points). If we were to use such a low volume fraction resolution and required the gradients to match exactly we wouldn't be able to identify two positions on the curve where the gradients were

[29]In equation 13.55 we include ΔG without any subscript to indicate the generality of this statement to the three different definitions for the free energy of mixing. In addition, we do not include the $k_b T$ factor solely out of convenience

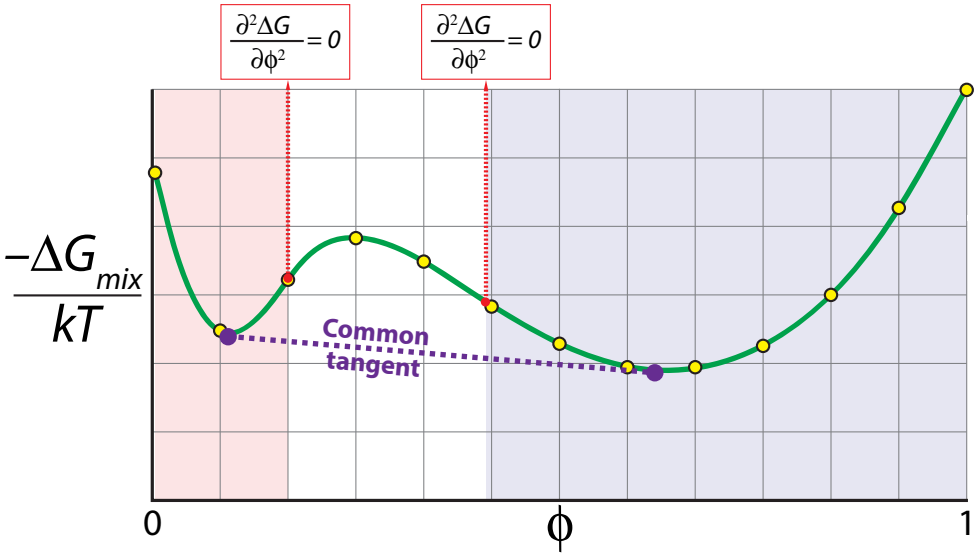

Figure 13.14: The free energy of mixing curve is shown in green, but we evaluate the gradient only at the yellow circles. The low concentration and high concentration search domains are highlighted in red and blue, respectively, as defined by the low and high concentration arm spinodals (red dashed line). Because our volume fraction resolution is so low, there are no two positions where the yellow circles evaluate to give a common tangent, despite the fact that there is clearly a common tangent line

the same, and may conclude that no common tangent exists on this curve. As is plain to see based on the purple line, a common tangent does exist, and while this is (clearly) an unnecessarily pathological example it highlights one of the challenges associated with identifying a common tangent. A second but related challenge stems from determining how similar two gradients need to be to conclude they are identical (i.e. what is the numerical tolerance). The numerical tolerance is intrinsically coupled to the volume fraction resolution - i.e. the lower the volume fraction resolution the higher the numerical tolerance should be.

As a final wrinkle, there is no guarantee a free energy diagram will have a common tangent, and the absence of a common tangent cannot necessarily be used to conclude the system has no two-phase regime. We already introduced an example where, despite the absence of a common tangent line, the two-phase regime is the energetically favourable configuration (see figure 13.12). In figure 13.15 we provide another example of a free energy of mixing curve where there is a well defined two-phase region on the free energy diagram, yet no common tangent. The minimum free energy tie line can here be defined as the line that minimizes the energy by running between $\phi = 0$ and tangent to the free energy curve without dissecting the curve at any point. It's worth pointing out that this type of free energy curve is - as far as we have observed - is *only* obtainable from the Muthukumar free energy of mixing. Such a curve *can* have two spinodal points (i.e. two places where $\left(\frac{\partial^2 \Delta G}{\partial \phi^2} = 0\right)$, but there may only be a high concentration spinodal. A sole high concentration arm spinodal corresponds to a region on the phase diagram where there is no instability region on the low concentration arm - instead the binodal and spinodal curves are collapsed on top of one another, meaning phase separation will uniformly occur through spinodal decomposition.

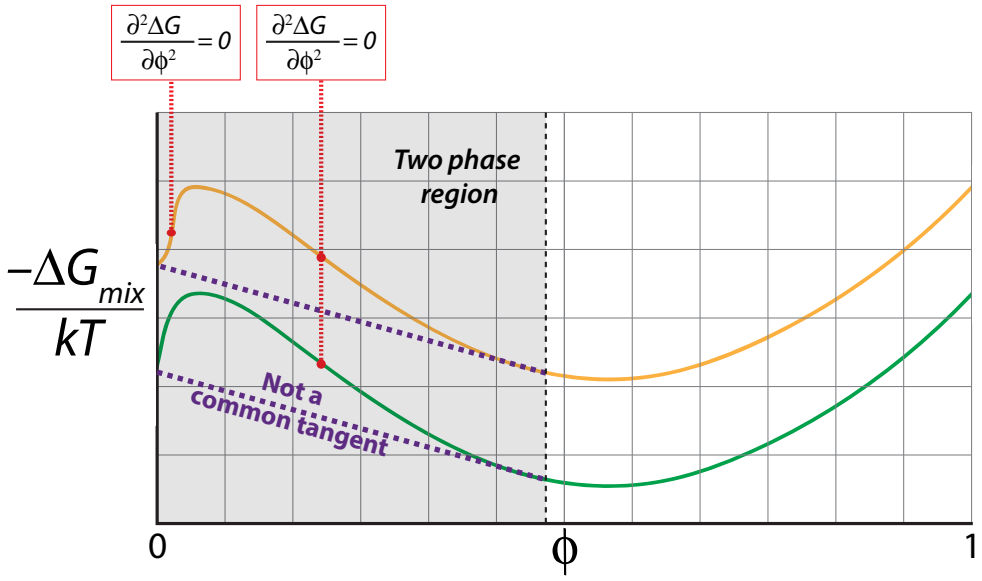

Figure 13.15: Two free energy curves that both lack a common tangent, yet contain a two-phase region (shaded in grey). Note the orange curve contains two spinodal points, while the green curve contains a single high concentration arm spinodal.

We have introduced these challenges explicitly to highlight the fact that while many literature sources make it seem as if constructing phase diagrams from free energy curves is trivial, there are a number of algorithmic, practical, and numerical challenges to overcome.

13.5 Fitting Muthukumar-Derived Phase Diagrams to Experimental Data: LAF-1

The preceding sections were (deliberately) fairly general, and provide a summary of the ideas, questions, and challenges associated with construction of phase diagrams. In the following two sections we focus specifically on the challenge of generating Muthukumar theory derived phase diagrams which are consistent with the experimental phase diagrams of LAF-1 and the LAF-1 RGG domain.

As a reminder, the Muthukumar free energy of mixing contains a number of parameters which are outlined below;

1. r - the degree of polymerization, which corresponds to the number of monomer units in the polymer in question
2. χ - defines the polymer-polymer vs. polymer-solvent balance (i.e., an effective two-body interaction term). This is proportional to the second virial coefficient
3. w - the three body interaction term (this is proportional to the third virial coefficient)
4. g_ξ - the concentration correlation length (mesh size) - informs on conformational fluctuations associated with the chain
5. α - the swelling ratio

kT is set to 1 as we normalize the free energy by kT.

Our goal is to identify the specific value for these parameters which allow us to reproduce the full phase diagram for the various LAF-1 constructs. In using the Muthukumar free energy

of mixing we make the assumption that we're working with a pseudo-binary system - i.e. changes manifest through salt and/or RNA will become evident in terms of changes to these parameters but are not considered explicitly. This is clearly a major oversimplification, but allows the theory to remain tractable.

The general approach used to determine the parameters which allow us to best reproduce the experimental phase diagrams is as follows:

1. Generate some χ vs ϕ phase diagram using a combination of parameters (varying χ) as described previously
2. Convert this phase diagram into a χ vs c (mg/ml) phase diagram
3. Compare the overlap between the experimentally derived phase diagram and the theoretically generated phase diagram
4. Update the parameters being searched in some way
5. Repeat until the best possible parameters are found

Thankfully we do not need to simultaneously solve for all six parameter outlined above. $kT = 1$ [30]. The degree of polymerization (r) is set to the number of amino acids in the constructs (708 for LAF-1, 168 for the RGG).

χ is extracted from the experimental data by converting the measured second virial coefficients (B_2) into χ using equation 13.56, where M_2 is the molecular mass of the polymer in question and $\underline{v}_1$ is the specific molar volume of a single unit on the Flory-Huggins lattice[31].

[30] Note that because we normalize by kT varying kT has no effect

[31] Specifically, this value is set to 0.018 liters / mol, based on the molar volume of water. This makes the simplifying assumption that water and solute monomers occupy the same space on the Flory-Huggins lattice,

$$\chi = \frac{1}{2} - \left(\frac{M_2}{r}\right)^2 \left(\frac{1}{\underline{v}_1}\right) B_2 \tag{13.56}$$

We make the simplifying assumption that α is salt independent and is equal to 1.0 (as discussed below this assumption appears valid). This assumption stipulates that the polymer does not substantially change its dimensions in the dense phase compared to the dilute phase. Is this reasonable? We argue that for LAF-1 it is; based on all atom simulations the RGG domain is already highly expanded, and the helicase-domains do not unfold in the context of the droplets. For the RGG domain to become *more* expanded in the context of the droplets would be associated with an enormous entropic penality. For the RGG domain to become more compact in the context of the droplet is inconsistent with its concentration. Taken together, we believe that the chain dimensions will not change significantly between the dilute and the droplet phase. Additionally, in a sensitivity analysis we found no material impact on the fitting procedure assuming $0.7 < \alpha < 1.2$.

g_ξ can be analytically determined by unpacking the Muthukumar theory. Specifically, the following is a key component of the theory;

$$\frac{1}{g_\xi^2} = \frac{6(1/2 - \chi + w\phi)\phi}{\left[\alpha^2 + (27/8)\pi(1/2 - \chi + w\phi) g_\xi \alpha^{-2}\right]} \tag{13.57}$$

This can be re-cast as a simple polynomial;

$$0 = g_\xi^2 - A_2 g_\xi - A_1 \tag{13.58}$$

and while we know this is not the case is a limitation of the Flory χ parameter which stipulates solute and solvent to be the same size

Where

$$A1 = \frac{\alpha^2}{6\phi\left(1/2 - \chi + \phi w\right)} \tag{13.59}$$

and

$$A2 = \frac{9}{16\pi\alpha^2\phi} \tag{13.60}$$

As a result, for each χ value we can solve the polynomial defined in equation 13.58 to obtain the two solutions associated with g_ξ. Given g_ξ must be a positive value, for valid combinations of the other parameters only one of these solutions is positive, which we take to be the value of g_ξ. The ϕ value here corresponds to the high concentration arm binodal point associated with the given χ value, which we are able to obtain from the experimental data - i.e., g_ξ is a parameter that is almost directly derived from the experimentally measured phase behaviour.

Finally, we are left only with w, which is the free parameter we are searching for in an unconstrained manner to identify the value of w which allows us to best reproduce the experimental data.

13.5.1 Comparing Experimental and Theoretical Phase Diagrams

Given the preceding section it should be clear that we need a quantitative way to compare the theoretical and experimental phase diagrams. The first challenge (which we need for obtaining the ϕ values for determining g_ξ) is how to convert volume fraction to concentration.

We estimate a constant factor ρ_0 based on the average volume of an amino acid. Assuming the average volume of an amino acid is 140Å^3 (1.4×10^{-25} liters), $\phi = 1.0$ corresponds to a molar concentration of 11.86 M. Assuming the average molecular mass of an amino acid is 110 g/mol, $\phi = 1.0$ corresponds to a mass concentration of ∼1310 mg/ml. Recall that

$$\phi = c\frac{1}{\rho_0} \tag{13.61}$$

Therefore,

$$\rho_0 = 1305 \text{ g/ml} \tag{13.62}$$

We use this as a general conversion factor to convert volume fraction to mass concentration. This is (without question) a simplification - notably we expect that the density (which ρ_0 is reporting on) is different between the dense and dilute concentration regimes (note that in this context density and concentration are not equivalent, clearly the concentration of protein in the dense phase is higher, but ρ_0 is describing the density of *all* material; solvent, protein, salts *etc.*). Consequently, we anticipate the need for a constant offset for the low and high concentration arms of the spinodal/binodal curves to account for these density offsets. ρ_0 should be considered a factor which allows us to convert ϕ into the right order of magnitude in terms of mass concentration, but not a universally correct constant.

Having established a general purpose way to convert between mass concentration and ϕ, we are left with another question: how can we quantitatively and efficiently compare the experimental and theoretical phase diagrams with one another?

The experimental phase diagrams for each construct contains six points (three on each of the low and high concentration arms) which define the coexistence boundary. We initially tried using these six points as constraints when performing searches for an optimal w parameter, whereby the theoretical value for ϕ at each χ corresponding to the experimentally observable value was compared with experimental value. Unfortunately, we found this lead to substantial degeneracy and was not a tight enough constraint. We also tried using the ratios of the two-phase regime at difference values of χ to generate a phase diagram specific set of unique ratio-widths. While this has the appealing property of being unit independent allowing for a direct comparison of volume fraction and mass concentration, it fails to capture the asymmetry associated with the low and high concentration arms, and so as an objective function does not correctly capture the key phase diagram elements (though comes close)[32].

To overcome this, we sought to determine an analytical fit to the experimental data, which would allow us to generate an arbitrarily large number of points along the coexistence curve for experimental-theoretical comparisons. We found that a rational polynomial fit was best able to capture the experimental data in an analytical form (see figure 13.16. Note that while this fit does an (apparently) excellent job of interpolating between the experimentally determined points, because it is solely a phenomenological fit there is no reason to assume it has any extrapolative power and should not be used to determine the coexistence curve beyond the lowest and highest concentration points.

[32]In this work we were required to perform the ϕ to concentration conversion to determine g_ξ. Had this not been the case, the widths ratio limitation in terms of the inability to capture the phase diagram asymmetry could be corrected for to give a general purpose approach for comparing volume fraction and concentration phase diagrams in a unit-less space. This may be an appealing route of investigation in the future for connecting experimental and theoretical work

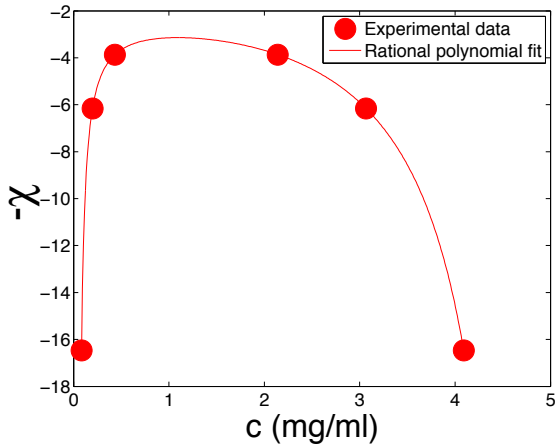

Figure 13.16: Rational polynomial fit to the experimental coexistence points associated with LAF-1 + 30k poly rA

13.5.2 Searching for the Optimal w Value

We have now defined a formal approach to compare experimental and theoretical phase diagrams in a robust and consistent approach. With this in hand, we searched through w space to determine the optimal value for the three-body interaction term. Initially we attempted to perform a global optimization of w, but found that the w goodness-of-fit surface was too rugged and non-continuous for standard optimization procedures. To overcome this difficulty we developed and deployed two entirely independent approaches for searching through w space.

A rapid local optimization search procedure involved randomly selecting some value of w and performing a short local optimization search to find the best value of w. This can be thought of as an approach for quickly survey the w landscape, which allows us to place some broad upper and lower boundaries on the possible values of w. Because w is proportional to the third virial coefficient we know it must be positive (i.e. lower boundary = 0). This local

optimization search placed an upper boundary on w of 200000. It is worth pointing out that the absolute value of w should not be considered a true three-body interaction value, but instead some $A \times w^{\text{true}}$ where A represents a constant units correction factor ($w = Aw^{\text{true}}$).

Having established these boundaries, we next performed a more computationally intensive Monte Carlo search to identify the globally optimum w value. Briefly, the search allows both local and global perturbations to w, where moves are accepted and rejected via the Metropolis Criterion. We ran sixty independent Monte Carlo searches for each construct and identified the single global optimal w for each one. The goodness of fit (lower values correspond to better fits) surface is shown in figure 13.17.

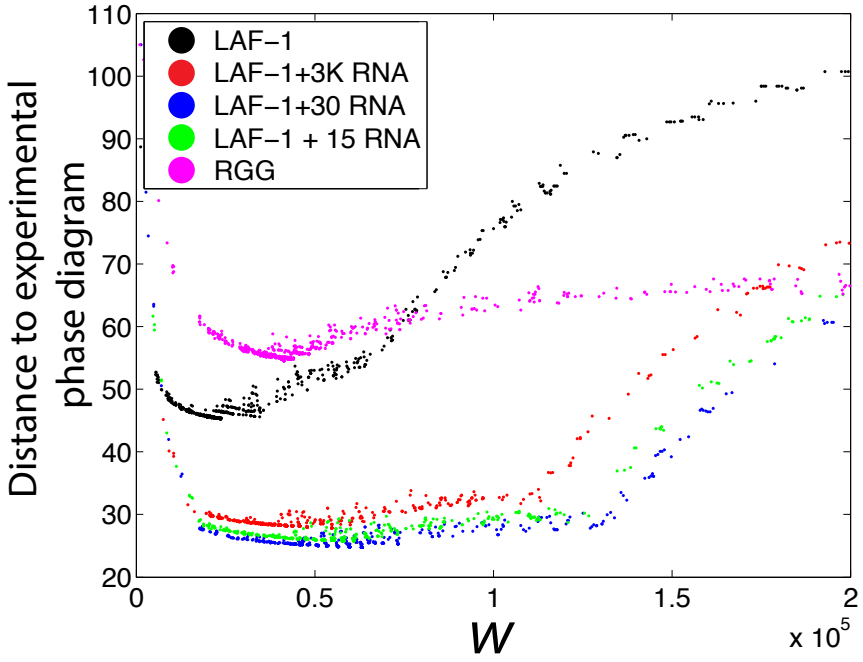

Figure 13.17: The goodness of fit for differe w values is shown from best fit results form 60 independent Monte Carlo simulations. Each point represents a single local minima identified during the Monte Carlo search procedure. Multiple minima are identified from each simulation.

Having performed an extensive Monte Carlo search procedure, we now have the construct and salt dependent g_ξ values and the construct specific w parameters which let us generate the best fit theoretical phase diagram. The raw best-fit diagram for LAF-1 is shown in figure 13.18a.

Despite getting the shape fundamentally correct, there is an apparent offset between the experimental and theoretical phase diagram on both the low and high concentration arms. It is worth noting that irrespective of the type of objective function used, we cannot make the Muthukumar derived phase diagram wider while maintaining the same profile in the χ dimension - i.e., this offset issue is not a limitation of our Monte Carlo objective function (importantly the raw Muthukumar derived phase diagram places the high-concentration binodal at slightly too dilute a value, due to the relative differences in density). As mentioned previously this offset is the anticipated error due to our assumption of a single constant ρ_0. We corrected for this with a low and high arm constant offset to correct for the density difference in the dense and dilute phase. The high-concentration arm density corrections should, in theory, be χ dependent, although in practice this χ dependence is extremely small so to a first approximation a fixed offset is entirely reasonable.

Figure 13.18b shows the offset-corrected comparison of experimental and theoretical phase diagrams using the optimal offset value. We define this offset by determining the uniform value that minimizes the difference between the experimental and theoretical high and low concentration arms (both arms are fit independently). The procedure generates a pair of offset values for each construct, and allows the theoretically generated phase diagram to recapitulate the behaviour of the experimentally derived phase diagram.

Having gone through this procedure, we are left with phase diagrams in χ vs. ϕ and χ vs. c space, as shown in figure 13.19

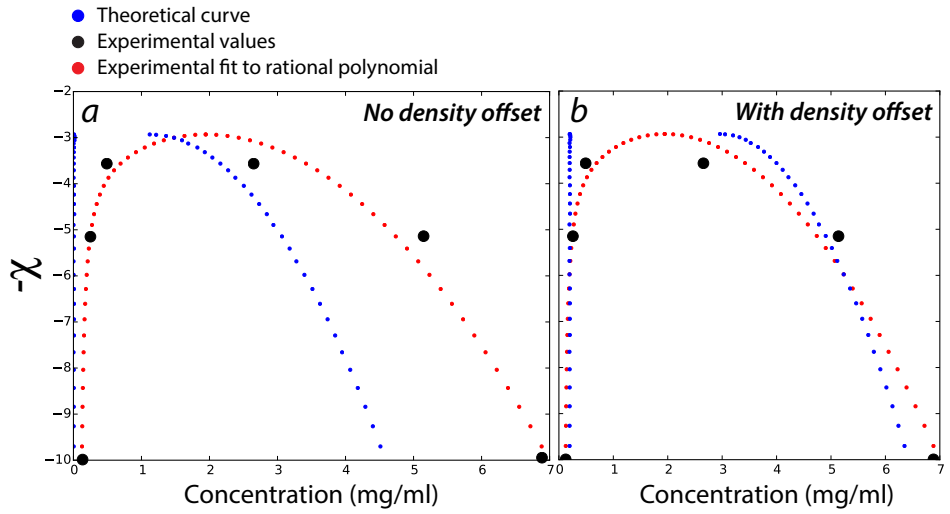

Figure 13.18: Comparison of the best LAF-1 theoretical phase diagram vs. experiment with (b) and without (a) the binodal density offset

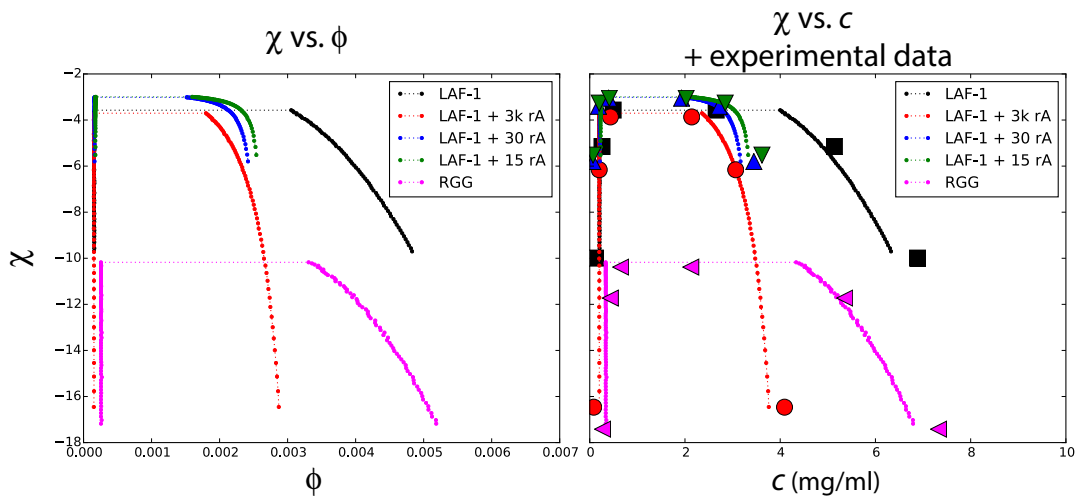

Figure 13.19: Full experimental phase diagram with theoretical fits shown in lines, cast in ϕ space and mass concentration space

13.5.3 Limitations of This Approach

While there are a number of theoretical limitations/caveats associated with the approach, we highlight a couple of numerical/procedural limitations worth considering.

The first is that by using the rational polynomial as a purely phenomenological fit we lack any predictive power regarding the phase diagram for the concentration range outside the values explicitly examined experimentally. This becomes an issue because our g_ξ value relies on the high concentration arm of the two phase regime (recall equation 13.58 contains a ϕ value), so extrapolating the theoretical curve beyond the values explored experimentally requires us to extrapolate how g_ξ behaves as χ becomes increasingly positive. Given the form of g_ξ as a function of χ (which appears to plateau out in the two phase regime - see figure 13.20) it may be possible to fit the g_ξ vs χ behaviour to an analytical expression to extrapolate to larger values of χ, although this is not an avenue we pursue in this work. This is a solvable problem, and more appropriate functional form would help alleviate this.

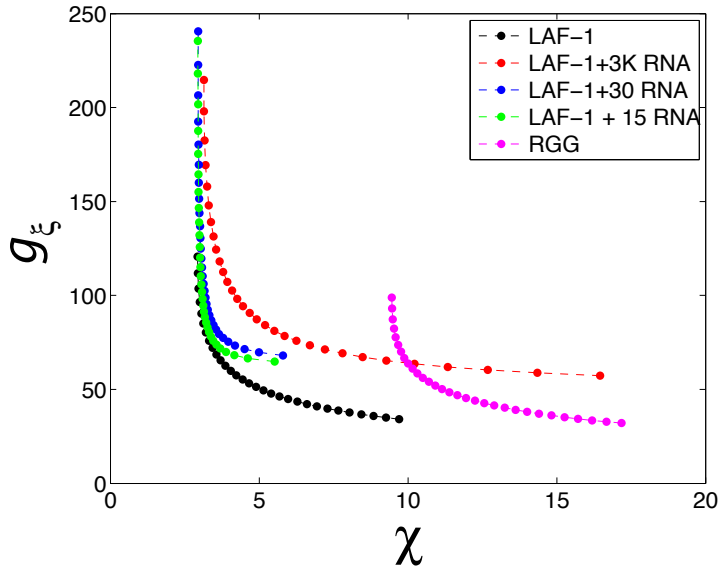

Figure 13.20: g_ξ vs χ - note that as χ becomes large g_ξ appears to be approaching a plateau, a result entirely expected as the two phase density increases (as salt decreases) to some maximum value

Secondly, by fitting the experimental data to a rational polynomial and using that as our objective function during the search process we are implicitly stating that the rational polynomial fit is correctly capturing the true phase behaviour. Given the error bars associated with the experimental high concentration values in particular, it is possible that the rational polynomial is imposing an unreasonable constraint onto the fitting procedure which creates an inherently unsatisfiable problem. That said, the key features we wish to capture from the experimental data (the relative widths of the two phase regime and the relative positions of the critical χ values) are well defined, such that even if specific numerical nuances are not quite correct the general principles that emerge from the derived w and g_ξ values are robust to such differences.

Thirdly, we treat the RNA as a mean-field perturbation to the free energy of mixing parameters, as opposed to as its own component in the theory. This is almost certainly incorrect; we are left with theoretical insights which correctly predict the mesh size in the presence and in the absence of RNA, but a more compelling description would be a variational method that is generalized to a ternary (or even n-ary) system. To develop such a theory is challenging. Binary mixtures are convenient from an analytical perspective as there can be only two coexisting phases. For a ternary (or higher order) system there are many more possibilities, as examined by Jacobs and Frenkel [256, 257]. While a correct treatment of RNA is necessary, it is unclear if mean-field analytical theories are an appropriate route to explore multi-component phase behaviour due to their inability to describe surface tensions between coexisting phases. Part of this concern is the motivation for our novel simulation engine PIMMS (see 14.

13.6 Discussion

13.6.1 The Decoupling of Intra- and Inter-molecular interactions

The key feature of Muthukumar's theory of polymer solutions that has allowed us to fit the LAF-1 data is the consideration of large-scale fluctuations. These fluctuations are on the same order of magnitude as the fluctuations observed in simulations of the RGG domain, suggesting that the theory is capturing fluctuations of direct relevance to the material properties of the dense phase. Beyond simply providing an analytical description of the density fluctuations within the system, the independence of ξ and χ allows Muthukumar's theory to decouple intermolecular and intramolecular interactions. In standard free energy of mixing theories the χ and w terms describe the interaction between monomers in an entirely mean-field manner; intramolecular solute-solute and intermolecular solute-solute interactions are not distinguished. This is entirely appropriate for a homopolymer, but for heteropolymers this is not necessarily the case. Muthukumar's theory allows a system to simultaneously have a large positive χ value (i.e. strongly attractive solute-solute interactions) yet simultaneously engage in large-scale conformational fluctuations. Such fluctuations implicitly require that intramolecular interactions are weakened enough to facilitate these fluctuations, leading to the an analytical description where the intermolecular interactions that are stronger the intramolecular interactions. This allows phase separation to occur at extremely low solute concentrations, but ensures that the volume fraction occupied by the chains in the dense phase remains large such that the concentration inside the droplets is very low. It is this decoupling that allows Muthukumar's theory to correctly capture the solution behaviour of LAF-1, likely via a 'stickers on a chain' style architecture as proposed by Semenov and Rubinstein; short short attractive motives distributed across the RGG domain sequence [532].

Given the necessary decoupling intermolecular and intramolecular interactions, can we find direct experimental evidence for this decoupling? usFCS allowed us to determine the diffusion coefficients at a range of salt concentrations (125 mM, 250 mM, 400 mM), which were used to determine the B_2. B_2 is a measure of the strength of the biomolecular interactions - it provides a formal description of a *by definition* intermolecular interactions. Simultaneously, using the measured diffusion coefficients (D) we can calculate radii of hydration (R_H) at a range of salt concentrations using the Stokes-Einstein equation (eq. 13.63) to ask how salt influences the intramolecular interactions.

$$D = \left(\frac{k_B T}{6\pi\eta R_H} \right) \tag{13.63}$$

Here, D is the measured diffusion constant, η is the solution viscosity and R_H is the hydrodynamic radius. Note that the the relationship between R_G and R_H is enormously dependent on the shape of the molecule (as well as additional and confounding effects such as the hydrodynamic effects of solvent slaving, polymer dynamics *etc.*) [68]. As D becomes larger (faster) R_H becomes lower (more compact). Concurrently, as intramolecular interactions become weaker, global dimensions increase (the chain interacts with itself less strongly) and the R_H becomes bigger. This provides us with the data we need to compare how - as a function of salt - intermolecular and intramolecular interactions change.

For convenience, we calculated these changes in terms of a normalized percentage to the value at 125 mM, and the results are shown in figure 13.21. B_2 changes by around 70%, while chain dimensions change by around 5%. These results provide a direct experimental demonstration of the decoupling between intramolecular and intermolecular interactions,

a behaviour necessary (although not sufficient) to explain the dilute nature of the LAF-1 droplets.

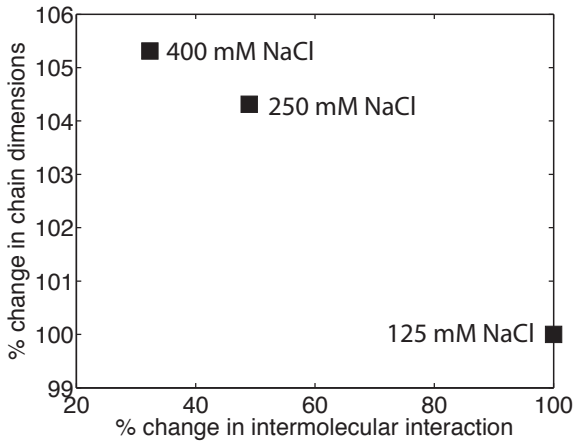

Figure 13.21: As salt concentration increases the intramolecular interactions weaken very slightly by $\sim$ 5%. In contrast, the intermolecular interactions weaken by around 70%. NaCl dramatically weakens bimolecular interactions with minimal effect of the individual chain's dimensions. These results are consistent with simulation results at different salt concentrations.

13.6.2 A Functional Role for Dilute Droplets

Why might the cell make dilute droplets? One possible answer stems from the fundamental metabolic cost of synthesizing high concentrations of proteins. As a thought experiment, let us work with the following estimates for some standard cellular values. The *in vitro* mass concentration of LAF-1 in droplets (with RNA) is ∼3 mg/ml under physiological salt concentrations. The volume of a *C. elegans* embryo is $\sim 467\,\mu m^3$, with about 50% of that volume accessible to soluble proteins [13]. With these numbers, we can simply ask what the copy number of LAF-1 needed to assemble a single droplet of an arbitrary concentration and diameter. This is shown in fig. 13.22.

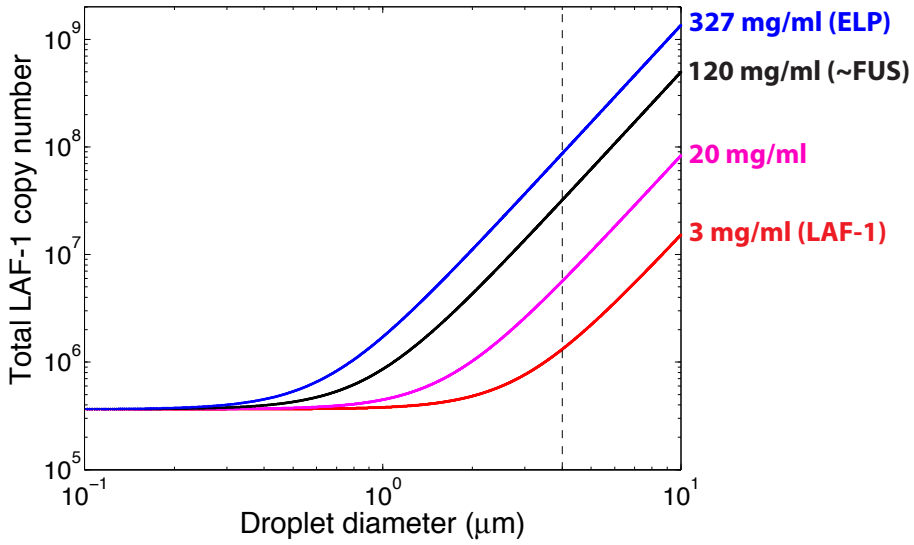

Figure 13.22: For dilute droplets the LAF-1 copy number is well within physiological protein copy numbers. However, for dense droplets the LAF-1 copy number rapidly exceeds the concentration of all proteins within the cell as droplet diameter grows.

For large dense droplets the metabolic cost of manufacturing enough protein to form the is intractable for the embryo. One could argue that P-granules are not just LAF-1 with many

other components contributing the droplet size. While this is true, for droplet with a diameter of µm2.5 and a concentration of 120 mg/ml requires $\sim 10^7$ proteins, suggesting a substantial fraction of the embryo's protein should be in a single P-granule. Embryos frequently have multiple P-granules, and P-granules are just one type of membraneless organelle. The burden of assembling 15-20 distinct droplets, each at a concentration of 300 mg/ml per droplet is simply not an energetically tractable route. We propose that dilute droplets, rather than being an unexpected result, may represent the only reasonable evolutionary path for the formation of large numbers of big droplets in the *C. elegans* embryo. This does not necessarily preclude the formation of denser droplets in smaller cells (the *C. elegans* embryo is significantly larger than most other cell types) but we speculate that for the formation of large droplets on the edge of phase separation (droplet formation and dissolution occurs in response to a gradient across a single cell) dilute droplets may represent an ideal mechanism for phase separation. A simple way to test this would be to identify a membraneless for which all the components (protein and RNA) have been identified and determine the cellular copy number of those components. This places an upper bound on the droplet concentration (assuming 100% of droplet components are found in the droplet). This is a simple experiment to describe, but a more challenging one to conduct.

A second explanation for the advantage of dilute droplets is based on the fact that dense droplets would be an inconvenient environment to facilitate complex biochemical behaviour. Within a dense droplet diffusion is necessarily slow, and the high concentration of protein would make passive diffusion into the droplet challenging without either an energy-coupled uptake mechanism or strong preferential interactions. A commonly stated explanation for the evolutionary advantages conferred by membraneless organelles is that they provide local bio-reactors, concentrating various components of complex biochemical pathways (e.g. RNA processing) into a single location to drive process efficiency. For this to be true the

components within the droplets must be accessible, otherwise the effective concentration is reduced by placing them inside droplets. We suggest that dilute droplets with a mesh-size above that of an 'average' folded protein provide an organizational strategy for simultaneously recruiting specific components while allowing passive diffusion of products out of these droplets.

Should this result be taken to mean that all droplets are dilute? Almost certainly not. In fact, we suspect that at ∼4-8 mg/ml the LAF-1 droplet may be among the most dilute. This is in part due to the fact that many P-granules must rapidly form and dissolve in response to the gradient of MEX-5 - the rapid formation of large droplets is most easily achieved by hyper-dilute droplets (in effect, the 'bang-for-buck' in terms of radial growth per-monomer is greatest when the droplet is at its most dilute). For numerous other proteins that are necessary and sufficient to drive the formation of *bona fide* membraneless organelles, a droplet concentration of 20-200 mg/ml is expected. In this context, we define membraneless organelles as condensates or quinary assemblies where a complex repertoire of additional species (e.g. protein and RNA) exist within the organelle; in effect, we believe the term 'organelles' should refer to micron-scale structures that can accommodate a wide range of additional clients. From a phenomenological standpoint, membraneless organelles are a natural cellular structure to identify and characterize, owing in part to the ability to visualize them using light microscopy. While many large membrane-less organelles have been identified, we predict that these may be the exception and not the rule, and that there will be *many* more assemblies on the sub-diffraction lengthscale (¡ 300 nm) that show many of the characteristics of a phase-separated condensate. Are these smaller condensates likely to be dilute? We suspect not. The sequence-requirements associated with disordered regions that form dilute droplets are expected to be much more restrictive than those for forming denser droplets, due to the need to decouple intramolecular and intermolecular interactions. The

need for dilute droplets is largely associated with scale and function. For smaller assemblies neither protein concentration nor encapsulation of large numbers of additional components is relevant, such that forming denser droplets seems much more likely. Consequently, we expect that for many assemblies, especially those driven by polar-rich IDRs devoid of charged residues, the intradroplet concentration will be around 200-300 mg/ml.

13.6.3 A Functional Role for Phase Separation in Biology

If not to act as organelles, why is phase separation a useful tool for biology? There are several properties of such a process that we believe make it attractive, as outline in 3. These are briefly recapped below in fig. 13.23. Phase separation provides an inherent **concentration buffering** effect, essentially offering energy independent proteostasis. Similarly, stimuli-responsive phase separation offers a binary mechanism to **sequester** soluble components. Condensates provide a general mechanism for the assembly of multicomponent **functional assemblies** and **signal integration** without strong evolutionary pressure for structure, although mediating specificity may be more challenging.

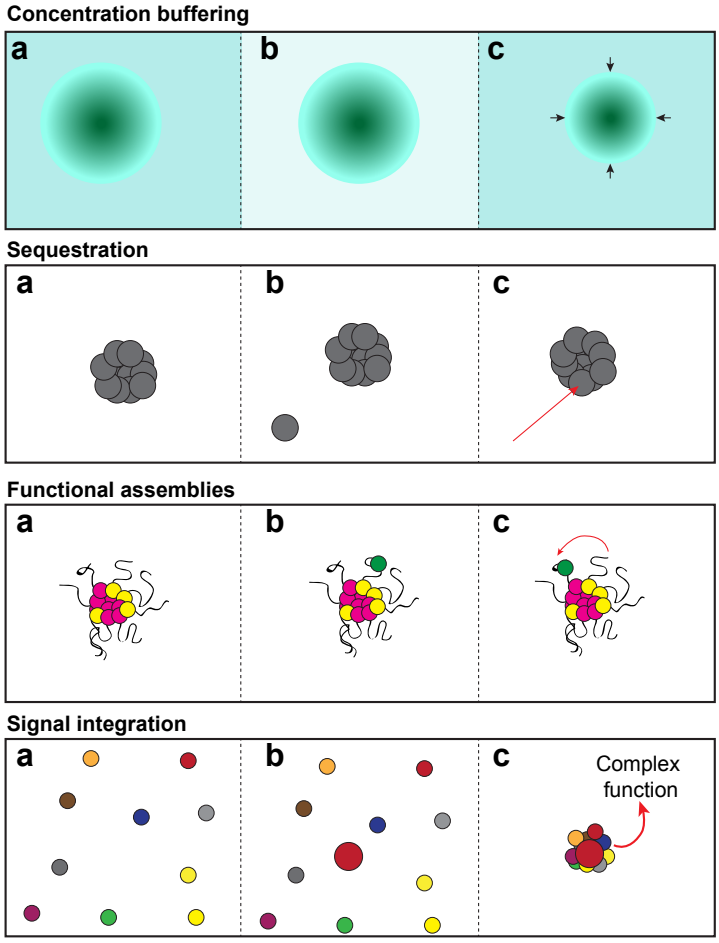

Figure 13.23: Putative droplet functions (see discussion in chapter 3)

13.6.4 Is Disorder Required for Dilute Droplets?

The requirement for dilute droplets, as defined in this work, is for the monomer to have a large pervaded volume, engaging in strong (and multivalent) intermolecular interactions without leading to chain compaction. These same design principles could be realized by linear species with sticky interaction sites. To this end, coiled-coils may be perfectly poised to undergo phase separation and form ultra dilute droplets (see fig. 13.24), although the balance between attractive patches, solubility, and non-specific repulsion would need to be

perfectly counterbalanced to avoid linear ordering and liquid crystal formation, as appears to occur in the synaptonemal complex [493].

In a similar vein on a different length-scale, amyloid-like fibrils could lead to an ultra-structure that despite having a high-concentration of fibres is primarily 'empty', (i.e. on some macroscopic length-scale, this would appear as a 'dilute' condensate). While this sounds somewhat abstract, this is largely consistent with the Balbiani body ultra-structure, and may represent a plausible strategy for forming large, hyper-stable meshwork assemblies [52].

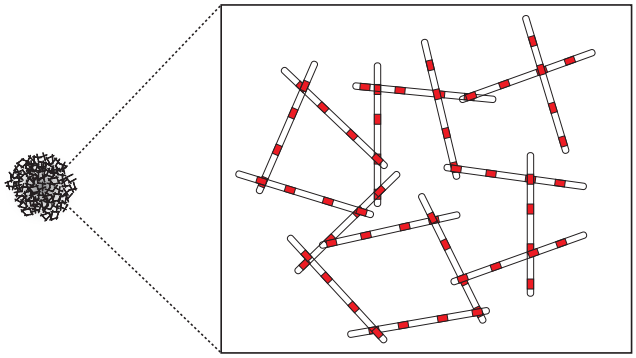

Figure 13.24: **Coiled Coils Could Form Dilute Droplets**. A schematic of a putative coiled-coil mediated meshwork facilitated by multiple weak 'sticky' patches distributed across the length of the coiled-coil domain.

Chapter 14

The PIMMS Simulation Engine

The following section is taken from a manuscript in preparation. All aspects the work were performed by A.S.H.

14.1 Background and motivation

The preceding sections provide ample examples of how the amino acid sequence of an IDP directly encodes its phase behaviour as a function of solution conditions, concentration, and temperature. Section I provides an extensive discussion on the mapping between amino acid sequence and conformational behaviour. Taken together, we can conclude that the amino acid sequence of an IDP will dictate the behaviour of both the individual chain's ensemble, and collective and emergent behaviour of multiple chains. As such we should - in principle - be able to predict and explore collective phenomenon as a function of amino acid sequence in much the same way as we and others have done in terms of the conformational behaviour associated with individual chains [125–127, 235, 359, 364, 405].

A major challenge associated with such a prediction is that this collective behaviour is, inherently, an emergent property of many disordered proteins together. While the development of novel theories that take sequence effects into consideration are in development, capturing the true chemical complexity presented by the repertoire of amino acids in conjunction with the highly irregular patterning of those chemical groups presents a major challenge for mean-field theoretical descriptions [335–337,517]. Moreover, the (typical) lack of three-dimensional representation associated with analytical theory presents a challenge in the decoupling of inter and intramolecular interactions. For homo-polymers this decoupling is at least partially (if not entirely) solved by Muthukumar's theory of polymer solutions, but the extension of this formalism to complex heteropolymers in n-ary systems (systems with many different heterpolymers) represents a major - a potentially unsolvable - challenge [256, 257, 409, 410].

Despite this, obtaining a predictive framework for mapping amino-acid sequence to phase behaviour would be extremely useful. An alternative approach is simulations. The size of the systems of interest make all-atom simulations impractical. For some sense of molecular scale, this could be thought of as folding 100-1000 proteins of $\sim 100-200$ residues simultaneously. This computational challenge could be solved by coarse-grained models (molecular dynamics, Langevin dynamics, Monte Carlo), but even the simplest models would like take weeks for a single simulations. Given the nature of the questions of interest, we would wish to perform a series of titrations in a concentration/temperature space to construct full phase diagrams, meaning that for even modest systems a minimum of $\sim$50 independent simulations (at unique temperature/concentration tuples) would be needed.

To address the challenge of computational cost and sequence specificity, we developed a novel Monte Carlo lattice-based simulation engine (Polymer Interactions in Multi-component MixtureS - PIMMS) and an associated amino acid-specific force field (the General Chemical

Forcefield - GCF). While both remain under development, and this chapter will not delve deeply into specific results, PIMMS allows us to perform sequence-specific simulations of hundreds of polymers to relative convergence on a single CPU in under a day. As a result of the underlying software architecture the computational cost-per-step scales as N with number of chains, allowing very-large systems to be simulated on a reasonable time-frame (hours to days). Converged single chain simulations take $<$ 1min, providing a high-throughput method for mapping sequence-to-ensemble relationships on a proteomic scale. Despite its simplicity, PIMMS allows interactions to be encoded over a hierarchy of ranges, allowing different types of interactions to be described in a way that recapitulates their behavior in high resolution models [67]. In the interest of clarity, we distinguish between PIMMS and GCF in much the same way that CAMPARI and ABSINTH are distinguished. PIMMS is a software package, and the vast majority of this chapter considers PIMMS in terms of simulations run with primarily phenomenological models to capture interesting physics. At the end of the results section, we will discuss early results from the development of GCF.

The remainder of this chapter is outline as follows. We introduce the model, how chains and interactions are represented, and what moves are performed, including a discussion of a new class of move for Monte Carlo simulations (Temperature Sweep Metropolis Monte Carlo). Next, we will qualitatively describe several examples of using PIMMS to explore collective chain behaviour. Finally, we will compare single-chain behaviour with results from all atom simulations and experiment to highlight the fact that, despite its simplicity, PIMMS + GCF are able to capture many sequence specific features with reasonable fidelity high fidelity.

We emphasize that PIMMS is still in development; while the vast majority of the simulation engine is complete, we are continuing to parameterize the associated forcefield to better reproduce all-atom simulations. Considering this the majority of the time associated

with PIMMS' development has been spent on developing an effective engine that balances physically relevant interactions with high performance and ease of use.

14.2 Methods

PIMMS is written in Python programming language. Analysis code and many of the more complex algorithms are written in native Python. High-performance components (notably the energy calculations and many of the underlying moves) are written in highly optimized Cython that compiles and runs at speeds comparable to native C. This provides us with a flexible programming framework with which to implement and test new ideas, but with the ability to re-code in a high-performance language as and when is necessary. Trajectory output is done using the MDTraj library [374]. Much of the code relies heavily on the numpy and scipy libraries. Version control is provided by GitHub. A website is available (http://pimms.xyz/).

14.2.1 The PIMMS model

PIMMS is an entirely generalized 2D and 3D lattice simulation engine. For convenience, we will present the basic principles in terms of 2D figures, but those ideas are identical in 3D.

Model overview

In PIMMS, the solute monomers are represent by beads, and the molecular units that undergo simulations are referred to as chains. Chains can be simple solutes (i.e. a single bead

per chain) or polymer of beads. We have developed PIMMS around the principle of representing IDPs as polymeric chains with a mapping of one amino acid to each bead, but in principle the mapping of molecule-to-chain could be any mapping of interest.

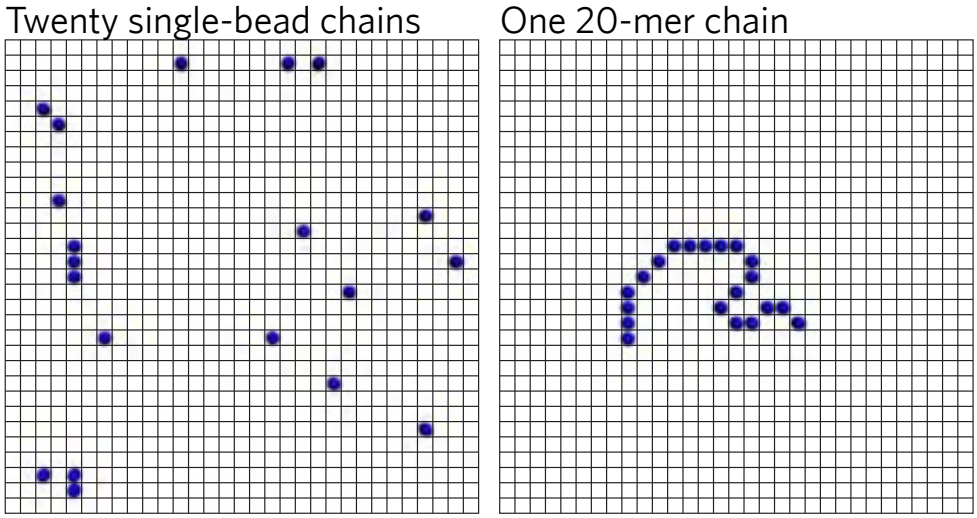

Figure 14.1: Possible configurations of PIMMS simple PIMMS chains

Beads have a fixed excluded volume; no two beads can occupy the same site on the lattice. Beads in polymeric chains are also constrained by chain connectivity; two beads that are consecutive in the amino acid sequence must be at adjacent sites to one another. Beyond this, beads are free to move to any site on the lattice. The lattice is a periodic environment; beads that pass through one face of the square (or cube) will re-appear on the adjacent side. Periodicity is entirely transparent to all molecules, such that all simulations are effectively done in an infinitely large box. We have developed several novel algorithms for the analysis of system spanning properties in a periodic space. The lattice configuration is updated via a set of Monte Carlo moves that maintain micro-reversibility (as discussed in subsection 14.2.2). If beads represent amino acids, then setting the bead-bead distance to be 4 Å provides a robust mapping between lattice dimensions and real-space dimensions.

Input

With the exception of the parameters being used, the simulation setup is defined entirely by a single input keyfile. This keyfile must also reference a parameter-file, which is used to define the relevant interactions (discussed below). For completeness, the full set of PIMMS keywords are described appendix A. The keyfile and parameter files are read by PIMMS in such a way that errors in setup information are explicitly reported on, as opposed to silently corrected. Extensive sanity checks are run during setup to ensure the combination of keywords makes sense, and the parameter-file is fully formatted without redundancy of re-definition of specific interactions. We felt that this was an important class of features to implement: we wanted to make it as difficult to possible for the user to accidentally run a simulation that would inherently not work or contained ambiguities with respect to important setup decisions.

Output

The lattice topology is written to a PDB file, and trajectories are written directly to the compressed XTC format. The trajectory can then be easily visualized using VMD [250]. All analysis output is generated as formatted text files, much of which is generated as the simulation runs, and a subset are generated as a summary at the end of the simulation. A wide range of analysis can be performed on-the-fly at an arbitrarily definable interval, meaning simulations generate the relevant analysis output *in situ*. Adding additional analysis routines is trivial, and can be loaded from free-standing code without modifying the PIMMS source code. Analysis routines include a range of analyses that are performed on individual chains (end-to-end distance, radius of gyration, asphericity, distance map, internal scaling),

cluster analysis routines that are performed on each topologically distinct cluster of chains (cluster size, density, asphericity, connectivity *etc.*), and general system spanning analysis.

Simulation Procedure

Upon the start of a simulation, all chains are randomly placed in non-overlapping configurations. If the lattice is dense leading to steric clashes between newly inserted chains, PIMMS will repeatedly try to place chains, but will eventually exit with an appropriate error warning. With all chains in place, the total energy of the system is then calculated, all pre-existing output files are deleted (if present) and the system prepares to enter into the Monte Carlo loop. Once the simulation begins a chain and/or move is randomly selected (based on the move frequencies defined in the keyfile) and performed, rejected or accepted according to the Metropolis Criterion, and the lattice configuration and system energy updated appropriately. At intervals defined by the keyfile, analyses are performed and output data is written to disk.

Equilibration

The equilibration period is a defined number of steps before analysis routines are activated. The number of equilibration steps must be less than the total number of steps, as equilibration steps use a subset of the total number of simulation steps. Equilibration can be standard, or can be performed as a thermal quench. The decision to quench during equilibration is set by the `QUENCH_*` keywords. In a quench equilibration, the system begins at some higher temperature and is gradually cooled to the production temperature. The rationale behind using a quench equilibration is as follows; given that during initialization all chains

are placed in a random non-overlapping conformation, the initial conditions could be considered as a snapshot taken from an ensemble at infinite temperature. As a result, the first few steps in the simulation are equivalent from an instantaneous temperature quench from ∞ to some very finite temperature. Such a quench could lead to the formation of local glassy states that become challenging to escape from - leading to significant challenges for sampling - and prevent a true equilibrium ensemble from being reacted. By performing a quench run, we allow this initial jump from infinite temperature to finite temperature to arrive at a high enough temperature that local glassy intermediates do not form. The production temperature can then be reached by gradually cooling the system, facilitating re-arrangement and reconfiguration to help avoid these locally stable states. Naturally the extent to which this help depends on both the quench start temperature and the rate of quenching, but nevertheless, for systems expected to undergo collapse it provides an effective approach to prevent (or reduce) the formation of meta-stable states.

Non-Bonded Interactions

PIMMS is based on a cubic lattice model, in which beads engage in interaction with all nearest neighbours. When site-adjacent beads experience an interaction energy, where the magnitude and sign of that interaction energy is defined by the parameter set. These interactions are referred to as short range interactions. Short range interactions can be positive (repulsive) or negative (attractive). In addition to bead-bead nearest neighbour interactions, all beads interact with empty sites according to a solvation energy. This allows us to capture hydrophobicity explicitly, instead of implicitly, and allows us to more readily encode complex chemical behaviour by decomposing bead-bead and bead-solvent interactions into two separate and independent energy terms.

To capture electrostatic interactions, we also have long range (LR) and super-long-range (SLR) interactions. Instead of nearest neighbour interactions, LR and SLR interactions experience interactions with beads at +2 and +3 sites, but will only interact with beads that have been defined as engaging in LR and SLR interactions. This allows us to directly encode a hierarchy of interactions into the model, and provides a mechanism to (at least qualitatively) capture the solution behavior of charged residues, which are almost always full solvated yet engage in long-range repulsive and attractive interactions.

Bonded Interactions

To capture the inherent stiffness of the peptide backbone, we encode an torsional angle potential defined in terms of the angle obtained between the three beads centered on a central bead of interest. Backbone angles are classified into one of three types (bent, crooked, extended),

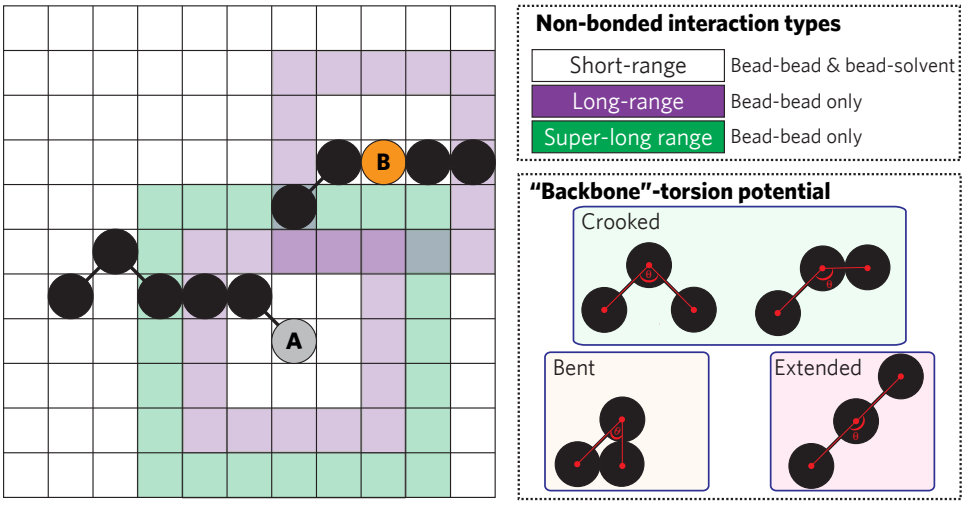

Figure 14.2: The components of the PIMMS Hamiltonian are illustrated. Beads experience non-bonded interactions (solvation, short-range [SR], long-range [LR], super-long range [SLR]) and torsional effects. In the figure, SR interactions are not shown in the interest of clarity, but every bead interacts with every adjacent site via either bead-bead interactions or bead-solvent interactions. The bead labelled **A** engages in LR and SLR interactions (as defined by the purple and green envelopes, respectively) while the bead labelled **B** engages in only LR interactions.

and a distinct angle penalty is associated with each. By parameterizing against residue-specific all-atom simulations, we are able to encode torsional angles that allow PIMMS to accurately reproduce all-atom simulations of repulsive Lennard-Jones terms on (i.e. the EV ensemble, see 2.4.4) in a sequence specific manner. Critically, this allows us to capture the flexibility of glycine and the stiffness of proline directly.

14.2.2 Moves

PIMMS employs a range of moves to aid in the exploration of conformational space. Below we briefly summarize the moves and how they perturb the system. All moves first determine if a steric clash has occurred (if a clash is detected the move is rejected automatically) and if not, the move is accepted or rejected according to the standard Metropolis criterion, depending on the change in energy associated with the move.

Crankshaft Move

Crankshaft moves are the main move-type that facilitate the evolution of single chains. The basic unit of a crankshaft move involves selecting the ith bead at random from a randomly selected chain, and moving that bead in a randomly selected X/Y/[Z] (Z if 3D) direction based on the relative positions of the (i-1) and (i+1) beads (i.e. such that the new position does not break chain connectivity). This new position is then accepted or rejected. This basic move is then used by randomly scanning through all the chains in the system and randomly selecting beads to move from each of those chains. As a result, a single crankshaft move leads to the complete update of all local positions. Due to the extremely local nature of the basic unit of crankshaft move, the entire procedure is incredibly efficient, meaning millions of crankshaft moves can be performed a minute.

Chain Translation or Rotation Move

Chain translation and rotation moves are two separate moves that involve randomly selecting a chain and performing rigid body translation or rotation in the X/Y/[Z] direction. Rotation

is either around 90/180/270 degrees to ensure the chain maintains the exact conformation (due to the lattice symmetry non-cardinal rotation results in relative positions of beads being moved).

Chain Pivot Move

Chain pivot moves involve randomly selecting a chain, randomly selection a position along the chain, and then pivoting either the long or the short half by 90, 180, or 270 degrees in a randomly selected direction. This allows large-scale conformational changes associated with a single chain to be rapidly achieved.

Chain Slither Move

Chain slither moves involve randomly selecting a chain, and having the chain 'slither' forwards or backwards in some random direction. These moves are relatively expensive and do not significantly change the conformational state of the chain, making them fairly ineffectual in the dilute and semidilute regime. However they become more relevant in dense systems.

Cluster Translate or Rotate Moves

Cluster translate and rotate moves are two separate moves where a cluster is selected at random and either translated or rotate around 90, 180, 270 degrees. A cluster is defined as a continuously connected system, where connectivity is defined as either part of a single connected network via short range interactions, OR a single connected network via short and long range interactions. In either case, depending on how the cluster was defined (i.e.

depending on what definition of connectivity was used), we do not allow a cluster move to merge two clusters together (as the reverse move would not be possible, breaking detailed balance). Cluster moves are important after an initial phase separation has occurred, as they allow discrete clusters of chains to move close enough to one another that coalescence can occur.

Single Chain Temperature Sweep Metropolis Monte Carlo (CTSMMC)

A single chain temperature sweep metropolis Monte Carlo (TSMMC) move involves randomly selecting a single chain and performing a series of local-chain crankshaft moves following the protocol outline in subsection 14.2.3. Once the full set of moves has been performed, the new configuration generated by the TSMMC moves is accepted or rejected, such that this can effectively be thought of as a single 'chain-rearrangement' move.

Multichain Temperature Sweep Metropolis Monte Carlo

A multichain TSMMC move uses the same procedure as described above, except a group of randomly selected chains are selected, instead of a single chain. With that difference, this move can be thought of a multi-chain rearrangement move. Again, only local crankshaft moves are performed on each chain. By default a max of 25% of the simulations chain's can be selected such that between 2 chains and 25% of the chains are randomly selected.

System Temperature Sweep Metropolis Monte Carlo

Finally, the system-based TSMMC move involves a TSMMC move performed across the entire system. For system TSMMC moves all other non-TSMMC moves are available. But no analysis our output is performed during the TSMMC cycle. At the end, the entire system reconfiguration is accepted or rejected. In effect, this can be considered as a complete system re-arrangement. Note that, TSMMC moves are effectively equivalent of running short exploratory simulations, and as such these moves are many orders of magnitude more expensive than the others.

14.2.3 Temperature Sweep Metropolis Monte Carlo

One of the advantages of Metropolis Monte Carlo (MMC) simulations when compared to molecular dynamics simulations is the ability of Monte Carlo simulations to escape local minima with moves that facilitate large-scale re-arrangements of phase space without breaking micro-reversibility (detailed balance). Various approaches have been taken to improve MMC sampling quality and enhance the ability to escape these local traps. These include Temperature and Hamiltonian Replica Exchange (T-REX, H-REX), parallel tempering, Hamiltonian Switch Metropolis Monte Carlo, and specific cluster moves to escape local kinetic traps [40, 196, 394, 501, 502, 565, 640, 656, 657, 667].

In the spirit of this, we have combined ideas from a number of different approaches and propose a new class of Monte Carlo move - the Temperature Sweep Metropolis Monte Carlo move (TSMMC). Unlike many of the other approaches, TSMMC is conceptually simple, trivial to implement into existing code bases, requires no parallelization, and makes no assumptions regarding the nature of trapped states. TSMMC maintains the detailed balance assumption via the incorporation of a correction factor to the moves' acceptance probability. For the derivation of this correction factor please see appendix B.

A graphical summary of the TSMMC procedure is shown in fig. 14.3.

The TSMMC move can be thought of as a single Monte Carlo move that can be selected and performed in much the same way as any other standard Monte Carlo move. The TSMMC move involves generating a series of auxiliary Markov chains at monotonically increasing and then monotonically decreasing temperature, where standard Monte Carlo moves are performed within those Markov chains at the various temperatures. In effect, it involves

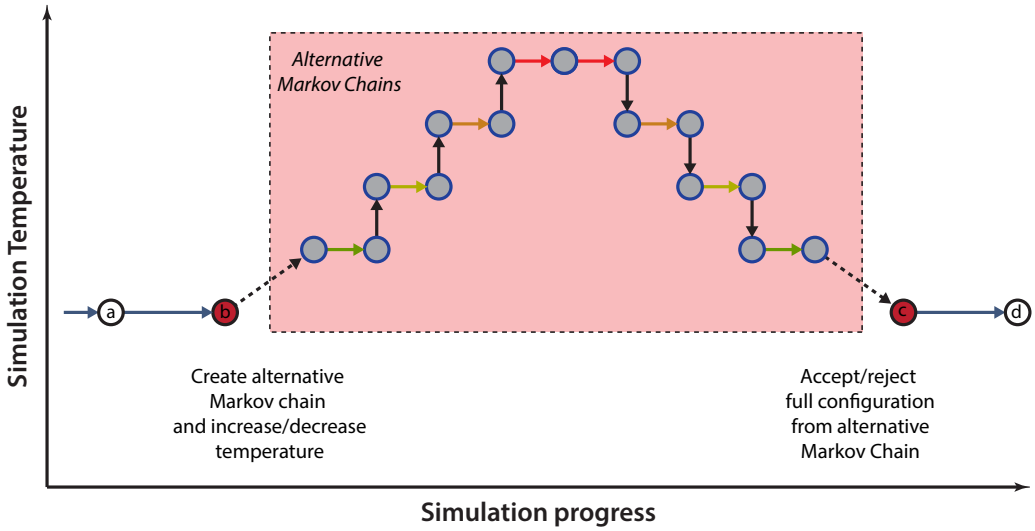

Figure 14.3: Graphical description of a TS-MMC move. Note that symmetry around the highest temperature is required to ensure the reverse transition path matches the forward transition path. Black arrows describe changes to the temperature and are rejection free, while coloured arrows represent a finite number of standard MMC steps at the given temperature.

gradually heating the simulation up and then subsequently cooling it back down in a symmetrical manner. Finally, the configuration obtained at the end of this heat/cool procedure is accepted or rejected. If the move is rejected the simulation assumes this new configuration and continues. If the move is rejected the simulation reverts back to the configuration immediately before the move began.

Figure 14.4 provides a graphical description of the approach. As a result, meta-stable kinetic traps can be escaped by these (effective) heating and cooling cycles, but the approach requires no *a priori* information regarding the behaviour of those meta-stable states or how to escape them. In this regard, TSMMC combines features of simulated tempering with the Hamiltonian Switch Monte Carlo of Mittal *et al.*, but is readily implementable in most

Monte Carlo simulation packages [394]. An important point to make is that if specific types of minima are frequently observed one can always design moves to aid in escaping these types of local states [640]. These moves will (likely) be extremely efficient, but require deep, system-specific knowledge to know what those states are. TSMMC abstracts the nature of local minima, instead allowing a conventional collection of Monte Carlo moves to allow natural evolution of the system as the relative depth of local minima are reduced.

As an example, we examined the behaviour of a generic homopolymer in a poor solvent undergoing phase separation with and without TSMMC, and reliably and reproducibly reached the final state in significantly fewer steps using TSMM (see fig. 14.4. A range of other tests showed robust improvements in simulation efficiency across a wide range of systems). Notably, TSMMC provides a mechanism for a system to exchange between deep but structurally similar energetic minima that are separated by large kinetic barriers. Importantly, we implemented a standard search Monte Carlo algorithm for the fitting of parameters on non-convex landscapes and found the inclusion of a TSMMC move robust enhanced the search process, suggesting that this provides a general, system agnostic approach to improve sampling with minimal overhead.

In summary, TSMMC provides an efficient and state agnostic approach to escape local minima, making it ideal for the enhancement of sampling in PIMMS simulation.

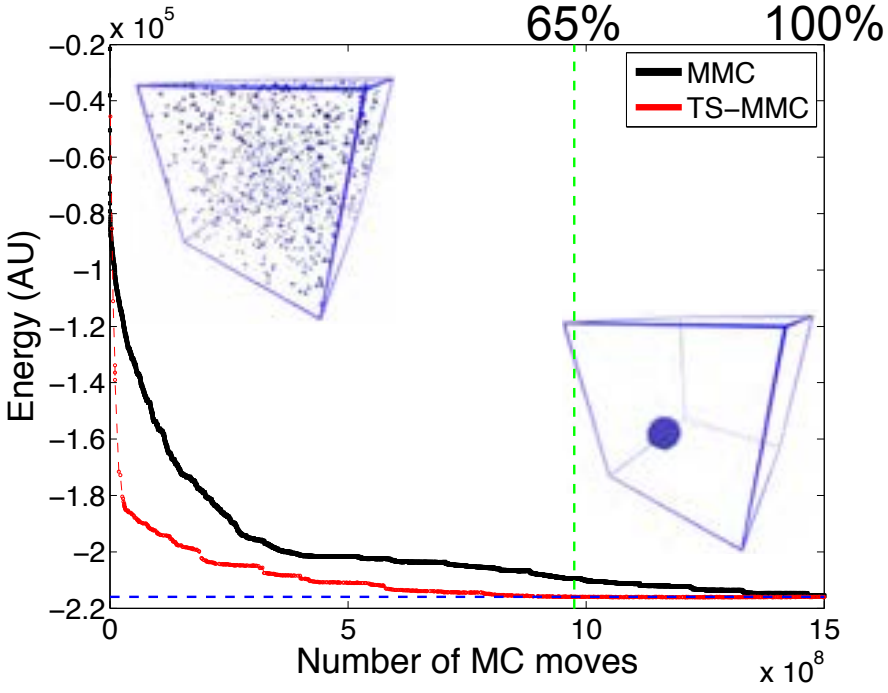

Figure 14.4: Number of steps vs. energy, whereby the global minimum is only achievable by a single spherical cluster. The TSMMC reached convergence in 65% of the time of non-TSMMC simulations. Traces report on the average behaviour taken from five independent simulations.

14.3 Results & Discussion

While PIMMS is still in development, we wish to provide a few vignettes of the types of questions we hope to be able to ask going forwards

14.3.1 Qualitative Phenomenologically-Derived Results

Design of Biologically Inspired Self-Assembly Materials

In collaboration with the Chilkoti and Zauscher groups, we are perusing the design of semi-synthetic partially ordered polymers (POPs) for self assembly into responsive materials for arbitrary functions [226, 329]. Preliminary data suggests PIMMS' ability to rapidly explore phase space in a high throughput manner and an ability to couple simulations to a machine-learning framework similar to the Gaussian Process Bayesian Optimization (GPBO) method pioneered by Ruff *et al.* allows for an automated approach to design specific types of polymer phases [506].

In polymer chemistry the need for iterative chemical reactions and cross-reactivity remains a barrier for the high-throughput design of chemically complex yet highly mono-disperse synthetic polymers. Methods developed by the Chilkoti lab allow for high throughput synthesis of highly repetitive amino acid sequences with high fidelity [469]. Coupling this process engineering with an ability to design material states based on coarse-grained sequence properties presents an opportunity to design and synthesize novel self-assembly materials with an arbitrary set of of functional constraints and features (temperature sensitivity, pH sensitivity, salt-sensitivity *etc*).

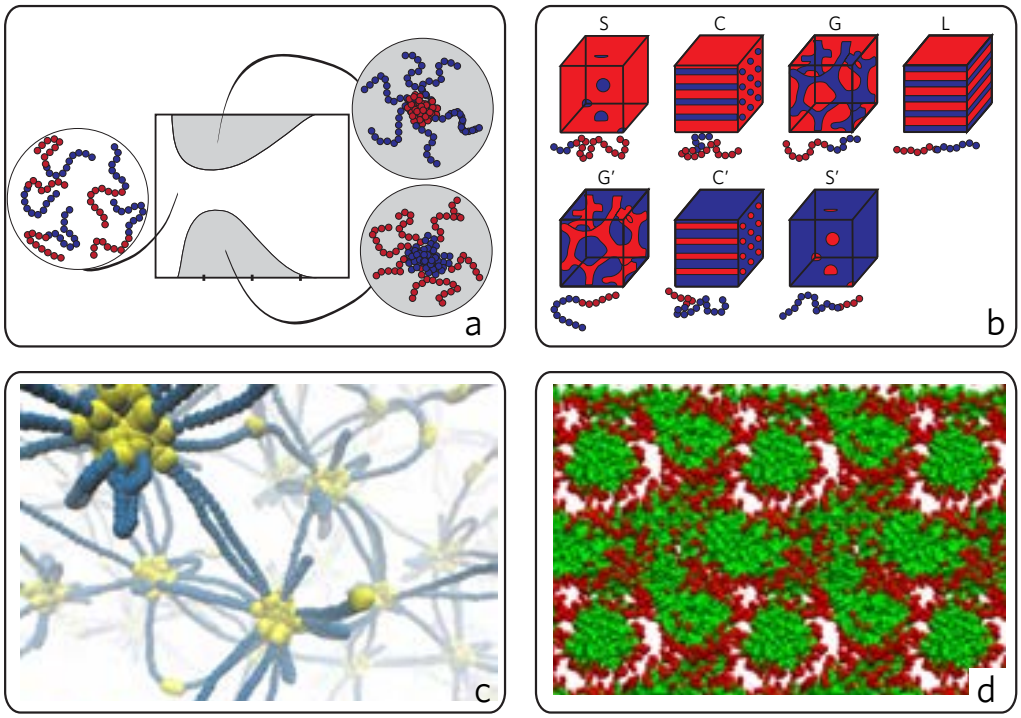

Figure 14.5: (a) UCST and LCST temperatures are tunable. Tuning can be achieved bashed on amino acid sequence, as described by Quiroz & Chilkoti and Roberts *et al.*, or via diblock copolymer block types [469, 489]. (b) In the synthetic polymer world a rich repertoire of phase space exists as defined by the relative strengths and sizes of diblock copolymers, an example of which is shown here (S = sphere, C = cylinder, G = gyroid, L = lattice) (based on figure from Khandpur *et al.* [286]). (c) PIMMS simulations of designed multi-block co-polymers form branched gels with well defined branch sites (d) At radically different volume fractions, similar synthetic sequences can assemble into complex semi-crystalline phases by modulation of interaction strengths and solubility of different blocks (3D simulation viewed from top down)

.

Determinants of Liquid-Liquid Droplet Mixing

An open question within the biological phase separation field pertains to the determinants of selectivity and mixing in droplets. The approach of reconstituting single components *in vitro* and characterizing their phase behaviour has been instrumental in exploring the sequence determinants of phase separation. However, our understanding of how compositionally distinct droplets can form, why they don't merge, and what allows droplets of many different components to form. Indeed, in theoretical work by Jacobs and Frenkel, a prediction from a random interaction models suggests that the expected behaviour for a n-ary system with randomly selected interaction strengths is to either form one, large, single condensate, or a small number of homogenous condensates [222, 257]. In cells, it *appears* that we see neither of these extremes. Instead we observe multiple heterogeneous condensates, although it is possible that in reality there are a small number of homogeneous condensates in terms of scaffold proteins (proteins that *drive* phase separation) with a heterogeneous collection of client proteins (proteins that partition into droplets but are not necessary for droplet formation). Never-the-less, there are several examples where a number of scaffold proteins are necessary and sufficient to drive droplet formation (e.g. PGL-3, LAF-1) raising question about the determinants of mixing [162, 218]. In unpublished data from the Alberti lab, several proteins which are known to co-localize and mix *in vivo* phase separate into liquid-like condensates *in vitro* but do not form mixed drops despite their adsorption on to one another.

Many of the proteins involved in phase separation contain RNA binding domains, and have IDRs that are able to interact with RNA. Surprisingly, for many of these proteins there appears to be little or no sequence specificity to the RNAs [33] [528]. We wondered if RNA could be play a role in driving the mixing of two distinct proteins. More generally, can two

[33] This is by no means always the case, see work from Zhang *et al.* [668]

polymers that phase separate independently be driven into a single, well mixed droplet by a mutually attractive third species. We performed PIMMS simulations to explore this hypothesis, and found that equivalently strong interactions with an 'RNA' polymer is necessary and sufficient to drive mixing of two previously immersible yet phase-separated liquids (see fig. 14.6). Although extremely preliminary, the broad RNA specificity and extremely high affinities ($K_D \approx$ nM to pM [162]) is entirely consistent with a model where RNA is a universal mixer. An alternative explanation consistent with these data is that if RNA is aggregation prone at the high cellular concentrations, these RNA-binding proteins may have evolved to form biomolecular condensates that act as RNA-chaperone assemblies, drawing RNA in, untangling and and linearising, and the spewing the RNA back out. More work must be done to explore the interplay between proteins and RNA, but all the evidence suggests this relationship is both complex and critical for the normal function of cellular condensates.

14.3.2 Solvent Mixtures Induce Re-Entrant Chain Behaviour

A polymer in a good solvent shows highly expanded conformational behaviour, as discussed in chapters 2, 5, 7 and various others. A curious result in the polymer chemistry world stemmed from the following observation: given a polymer and two distinct but miscible solvents that individually act as good solvents for the polymer in question, the titration of one of those solvents into the other leads to first order polymer collapse, followed by gradual expansion back to the good solvent limit. This behaviour cannot be captured by mean field theories; by definition, the solvent quality remains 'good' through the titration, yet robust chain collapse is observed to a poor solvent limit.

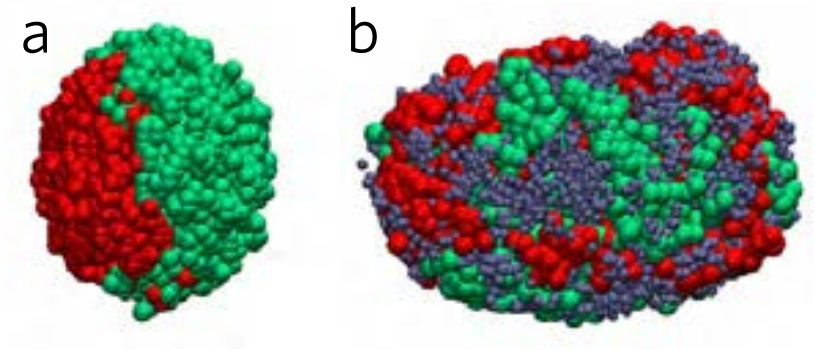

Figure 14.6: Panels a and b show the same phase-separated droplet with and without the addition of 'RNA' (grey polymer), a polymer with equivalent affinity for both the green and the red polymer. Upon addition of RNA the blue and green polymers are able to mix, and the RNA, acts to facilitate this demixing. This originates from a combined effect of bridging interactions and an entropic components associated with mixing within the droplet; effectively the RNA acts as a good solvent for both chains, making them miscible with one another in the process.

In 2014 this conundrum was solved by Mukherji, Marques amd Kremer, who showed that mean-field theories were indeed unable to capture this behaviour, which was a result of solvent molecules of one type bridging chain-chain interactions due to preferential interactions with the chain over the other solvent [403]. To use a consistent nomenclature for the remainder of this subsection, if we designate the polymer P and the solvents S_A and S_B, both S_A and S_B can have preferential interactions with P that out compete P-P interactions. However if one solvent has marginally stronger preferential interactions with P than the other then upon addition of that stronger binding solvent polymer collapse can be observed as the polymer maximizes interactions with the stronger solvent. In effect this phenomenon is equivalent to ternary hydrophobic effect *in trans*. Reports of two experimental systems display this behaviour are shown in figure 14.7.

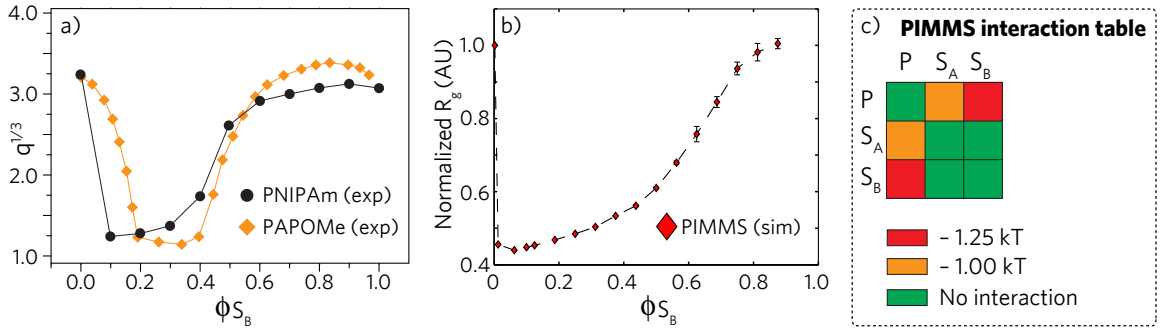

Figure 14.7: (a) Experimental PNIPAm data from Walter *et. al* and PAPOMe data from Hiroki *et al.*, the swelling ratio is directly proportional to radius of gyration, and describes the re-entrant behaviour in normalized units [231, 623]. (b) PIMMS simulations reproduce the first and second order transitions, as well as the sigmoidal shape of the second order transition as ϕS_B increases. Error bars are standard error of the mean between different independent simulations, and reflect the fact that PIMMS simulations provide exceptional sampling.

To determine if PIMMS is able to capture this behaviour, we ran simulations with a simple phenomenological model. A polymer of $n = 20$ was used for convenience, and the interaction table associated with the simulation is shown in 14.7 (recall that negative interactions are attractive). PIMMS correctly captures this re-entrant behaviour, but in addition captures the cooperative nature associated with the sigmoidal second order transition. The error bars are substantially smaller than in previously published simulation work [403]. It is worth noting that we are *not* trying to reproduce the exact behaviour observed here (in terms of the ϕS_B at which the first order transition happens or the specific steepness of the second order transition) - these are details that originate from the strength of the specific interactions associated with these two experimental systems. However, we sought to capture the general trends observed, in terms of the shapes of both transitions, which we have done

effectively. While this does not represent a novel result, it demonstrates how simply, quickly, and robustly PIMMS is able to directly tackle generic questions in polymer-physics.

14.3.3 Charge Patterning Modulation of Complex Coacervation is Generic

In chapter 11 we examined the phenomenon of protein-mediated complex coacervation. Complex coacervation describes a phase behaviour whereby two species are individually soluble, but upon mixing they 'complex' together, lead to the formation of a new solute-rich phase that contains both components [548]. In our work on NICD, we determined that the extent of charge patterning could be used to tune the driving forces for phase separation. We wondered if these result was specific to NICD (a related result observed in the N-terminal domain of Ddx4 suggested not), or a more general principle [421].

To test this hypothesis, we used PIMMS simulations to explore the phase behaviour of three simple synthetic designs for polyelectrolytes that titrate the degree of charge patterning. The three designs are shown in fig. 14.8. For each of three systems the composition of beads is held fixed, as is the patterning associated with the polyanion (red-bead containing polymer). However, the degree of patterning is titrated between three extremes.

To assess how these three patterning designs would influence phase separation we ran simulations with a fixed concentration of polyanion and variable concentrations of polycation. These were run at a range of temperatures. The concentration of polycation (x-axis of plots in figure 14.8b) is described in terms of relative concentration of polyanion units. All simulations have 100 separate polyanion chains, such that at a [polycation] of 1.0 there would be an equal number of polycation chains. In agreement with NICD results, we find that while

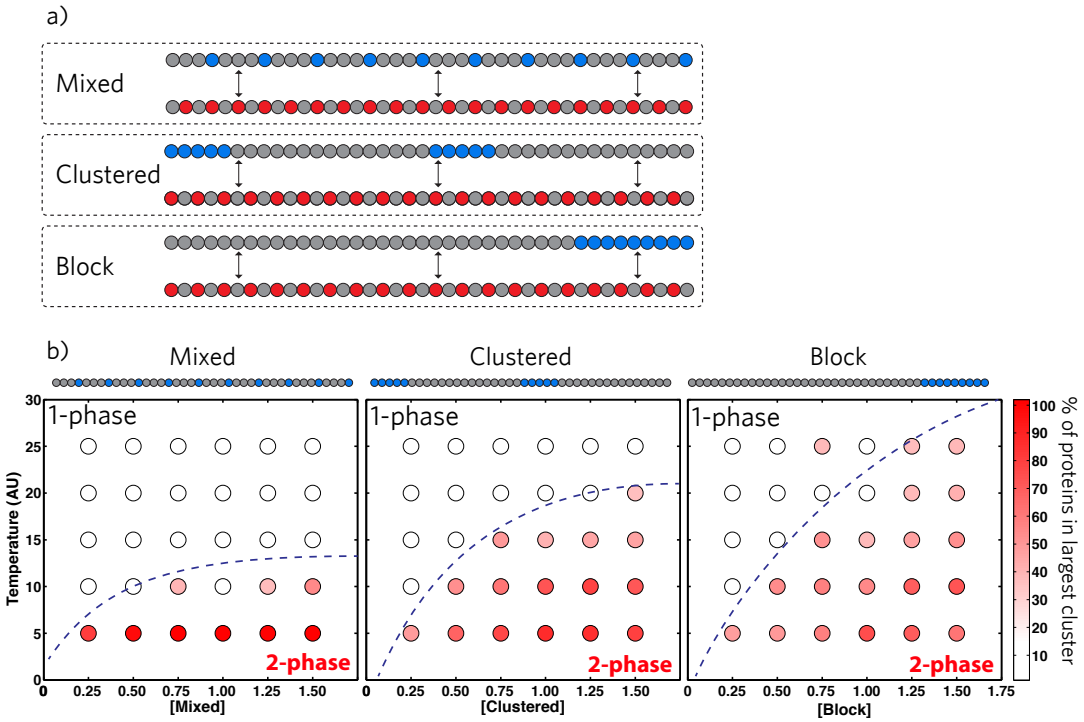

Figure 14.8: (a) Three distinct systems where the polycation (polymer with blue beads) has different degrees of charge patterning. The polyanion is held fixed a well mixed and more charge dense sequence. (b) Phase diagrams constructed from the three systems. The driving force for phase separation is weakest for the well mixed sequence, intermediate for the clustered sequence, and strongest for the block sequence.

the well mixed sequence has a relatively low propensity to phase separate the clustered and block sequences show a strong driving force to form single large clusters, which we take as a proxy for phase separation.

This result is consistent with recent work from Lin & Chan, who via a mean field theoretical treatment (random phase approximation) arrive at the same result [335]. The block vs. clustered assemblies show very different morphologies and internal re-arrangement, with

assemblies formed through block-based coacervates appearing more 'solid' like. While extrapolating kinetic properties from Monte Carlo simulations is a dangerous game, we tentatively suggest that the large charge blocks are the more dense clusters will be, though this will not *necessarily* mean a corresponding increase in viscosity. Further work is required to unpack this result, and additional components (salt, sequence length, spacing of clusters normalized by chain length) are being investigated to construct a general framework for describing the role of sequence patterning on complex coacervation.

14.3.4 Lattice Models can Capture Residue-Specific Local Behaviour

The preceding examples have focused primarily on phenomenological polymer models to reproduce generic polymer effects. Is a lattice-based Monte Carlo simulation engine able to capture details of relevance to intrinsically disordered proteins? We concede that the extreme simplicity associated with PIMMS interaction modes and geometry make capture much of what occurs in terms of local protein structure (notably secondary structure and folded structure) unobtainable. However, we also believe that PIMMS is robust enough to provide general coarse-grained properties regarding sequence-to-ensemble relationships, as well as provide insight into emergent phase behaviour in a sequence specific way.

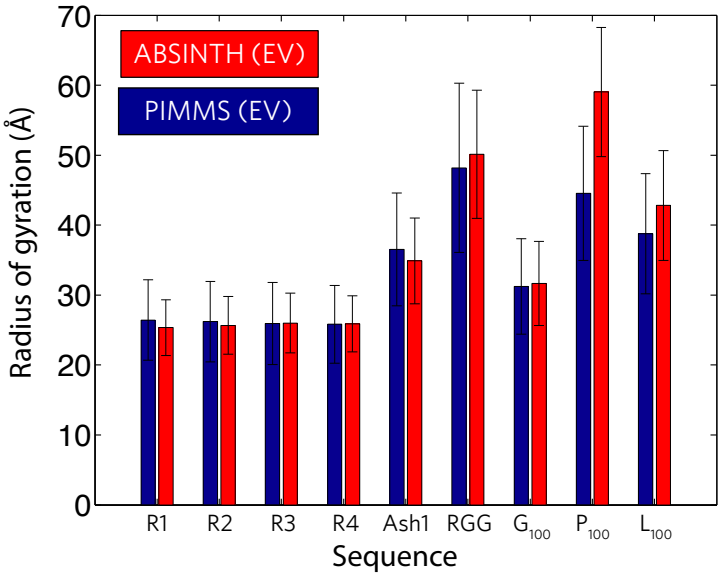

Figure 14.9: We compared EV simulations from ABSINTH with EV simulations from PIMMS to assess how robustly the PIMMS backbone potential can capture inherent sequence-specific conformational behaviour as a function of steric interactions.

To achieve this goal we must demonstrate some degree of agreement with experiment and/or higher resolution (and accurate) simulations. To this end, we first sought to reproduce sequence-specific chain dimensions obtained at all-atom resolution using an excluded volume (EV) Hamiltonian ($\mathcal{H}_{\mathrm{EV}}$, see subsection 2.4.4). We sought to use ABSINTH^{EV} simulations to parameterize the GCF torsional potential such that we could (hopefully) generate PIMMS^{EV} ensembles that match ABSINTH^{EV} ensembles in a sequence specific manner. The designation ^{EV} here refers to the fact the non-bonded interactions are solely describing the excluded volume effect, but the inherent local structure is still maintained (in ABSINTH via the all-atom model, in PIMMS via the sequence-specific torsional potential).

We ran all-atom ABSINTH^{EV} simulations of homopolymers of each of the twenty amino acids, and then generated 2D normalized histograms of the radius of gyration (size) and asphericity (shape) associated with those simulations. We then ran a full parameter sweep of angle penalty values for the bent and crooked angle classes described in fig. 14.2 on equivalently long polymers with PIMMS^{EV}. Finally we matched the re-normalized all-atom and PIMMS simulation distributions to determine residue-specific angle penalties. As well as correctly reproducing the homopolymer behaviour, the best fitting torsional potential parameters made intuitive sense; proline and glycine were the most and least restrictive respectively, and appropriate degrees of flexibility matched the expected sidechain size associated with each of the amino acids.

To verify if these parameters worked in normal sequences, we examined sets of polypeptides with randomly selected amino acids (R1-R4) as well as a several IDPs that we have worked with previously (Ash1 and the RGG domain of LAF-1) and three homopolymeric sequences, and compare PIMMS^{EV} simulations (using our new torsional potential) with ABSINTH^{EV} simulations. The results are shown in figure 14.9. PIMMS does well for the the sequences

tested here, although there is a modest under-estimation of the P_{100} radius of gyration. We suspect this originates from the angle potentials matching both size and shape, such that this trade-off ensures the ensemble asphericities are closer than if the fitted torsional potentials had been optimized solely to match the radii of gyration. However, the relative impact of glycine, proline and leucine are captured correctly, and in proline and glycine rich sequences (Ash1 and RGG) PIMMS^{EV} correctly reproduces the ABSINTH^{EV} radius if gyration. Taken together, these results give us confidence that, to within some degree, sequence-specificity is achievable with a lattice model.

14.3.5 Sequence-Specific Effects Induced by Charge Patterning

Having established that we can capture residue specific backbone behaviour in the EV limit, we sought to determine or ability to reproduce sequence specific behaviour in terms of charge patterning. The internal scaling profiles associated with the EK peptides of Das and Pappu provide an ideal test case [126]. In this work, it was shown using all-atom simulations that for IDPs of identical amino acid composition radically different conformational behaviour can be obtained as a function of the degree of charge patterning. This patterning is quantified by the parameter κ. For a *lengthy* discussion on charge patterning and the parameter κ see chapter 4.

We took the set of sequence used in the study by Das and Pappu and ran PIMMS simulations with our *in development* GCF forcefield. For reference, the all-atom simulations run in the 2013 study took $\sim$2 week to run, whereas the equivalent PIMMS simulations took $\sim$2 minutes. A subset of the internal scaling analysis is shown in fig. 14.10 (for a description of how to read internal scaling profiles please see chapter 5). PIMMS was able to accurately

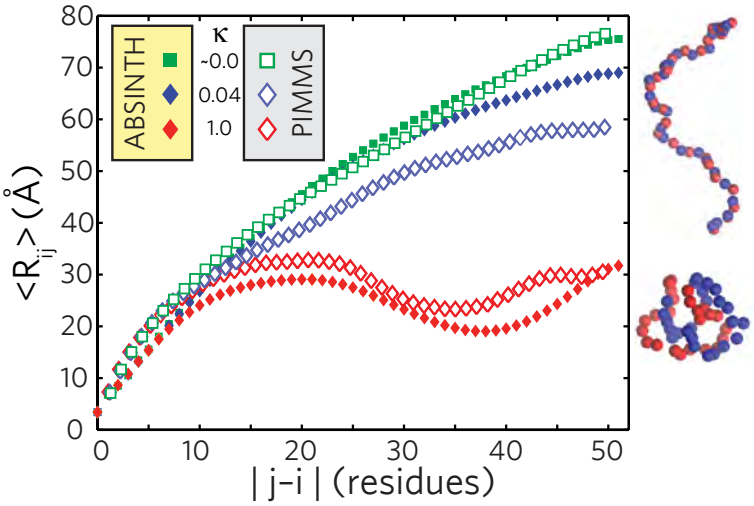

Figure 14.10: PIMMS simulations are able to accurately reproduce local conformational behavoiur of EK sequences, matching results from Das & Pappu [126]. The conformations to the right are representative conformations taken from the PIMMS simulations.

reproduce sequence-specific effects for all thirty sequences, although for clarity we show the most extreme variants here. These results suggests that relatively simple models can capture intricate and emergent properties determined by sequence patterning on a local level.

14.3.6 Sequence-Specific Radii of Gyrations from IDPs

The conformational behaviour associated with sequence specific effects described in the preceding subsection are promising, but disordered proteins with an FCR of 1.0 (i.e. every residue is charged) are rarely found in nature. Can we develop a more general, transferable parameter-set (a 'forcefield', although we use this term loosely) to capture residue-specific bonded and non-bonded interactions on a lattice? This model, which we call the General

Chemical Forcefield (GCF), is very much still in development, but in the interest of completeness we report early results here. Briefly, the forcefield has been developed using relative interaction strengths and solvation free energies taken from the ABSINTH model, with additional corrections made to account for the differences in residue size and structure. These parameters have been iteratively tuned against ABSINTH and experimental data from a variety of sources. Amino acids are divided into a set of 13 different classes to dictate non-bonded interactions, while each residue has its own independent solvation interaction (see table 14.3.6. Mixing rules (the non-bonded interactions between different groups) are defined on a per-case basis, as opposed to standard mixing rules used for Lennard-Jones parameters. The combination of bespoke and independent pairwise interactions coupled with a separate solvation term allows for fine tuning within a class to be dictated by residue-solvent interactions. The non-bonded, backbone, and solvation values associated with each amino acid have physically intuitive interpretations.

Can PIMMS + GCF reproduce sequence-specific radii of gyration? We asked this question with a set of fifteen well studied IDPs. We specifically chose sequences where well defined experimental data exists, ones were we had extensive simulation data, or both. Proteins where only all-atom simulation data are available have a * associated with them in figure 14.11. For many of these proteins, standard forcefields have been unable to accurately capture sequence specific behaviour. Sequences vary in length, and are between 15 (G_{15}) and 236 (both DDX4N1 and EGFR) residues. Simulations ran for $\sim 1 \times 10^8$ Monte Carlo steps with the first $\sim 1 \times 10^7$ discarded as equilibration. Simulations take between 2 and 5 minutes to run. A comparison of PIMMS-derived radii of gyration and experimental values is shown in fig. 14.11.

Residue group	Associated Amino Acids
Hydrophobic	I, L, M, V
Aromatic	Y, F, W
Proline	P
Glycine	G
Alanine	A
Arginine	R
Lysine	K
Negative	E K
Weak Polar	C S
Histadine	H
Threonine	T
Strong Polar	Q N

Table 14.1: Definition of residue groups in the current implementation of GCF (GCF_7E

Experimental data for the proteins is as follows. The RGG domain of LAF-1 (residues 1-168) has been studied extensively, most recently by simulation and FCS in chapter 12, although no independent measure of R_g has been performed. The N1 domain of Ddx4 (residues 1-236 - DDX4N1) has been studied extensively, and the R_g obtained here is from all-atom simulations which are in qualitative agreement with R_H measurements made by PFG-NMR [421]. α-synuclein has been extensively studied by many techniques, and the radius of gyration here is approximately equal to the average value reported throughout the literature [526]. CTD refers to a region of the RNA polymerase II C-terminal domain that

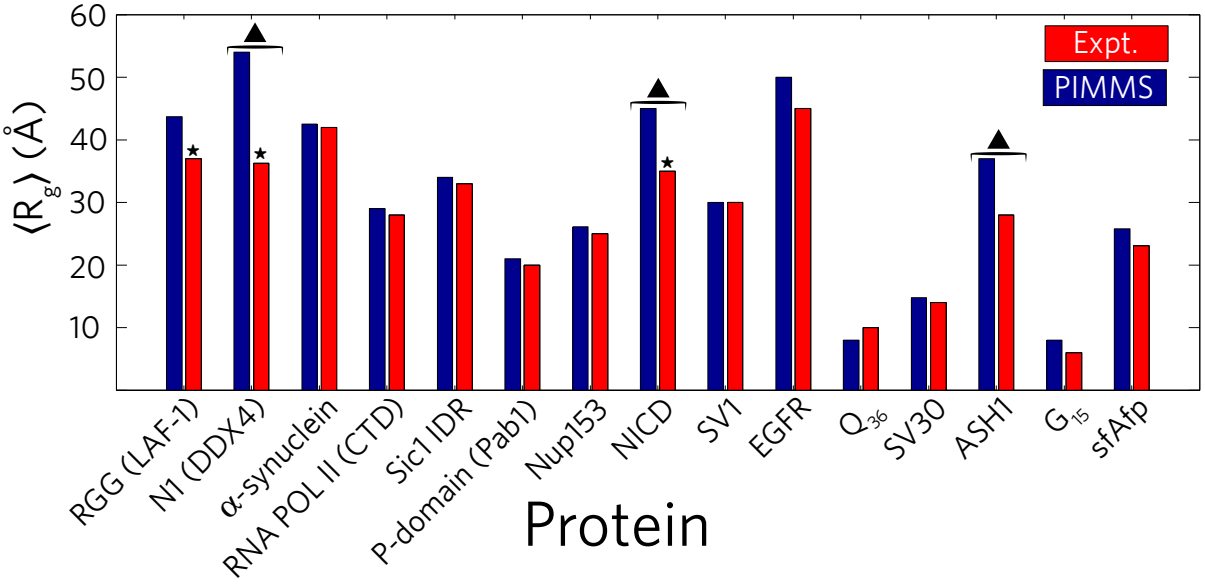

Figure 14.11: Comparison of experimentally and/or computationally determined radii of gyration with PIMMS-derived radii of gyration.

has been characterized by SAXS[34]. Sic1 (residues 1-90) has been studied extensively and shown to be highly expanded [35] [392, 393]. The P-domain of Pab1 (residues 419 - 502) was extensively characterized by SAXS in recent work [483]. Nup153 has been characterized by SAXS, smFRET and extensive simulations [378, 386]. NICD has been studied in the context of complex coacervation (see chapter 11, and while experimental structural data is lacking, all atom simulations correctly predicted emergent behaviour. SV1 and SV30 have been characterized by simulations and represent the two extreme κ permutants described by Das & Pappu [126]. The EGFR tail (EGFR isoform 1 precursor residues 718 to 943) has been characterized by simulations, but accurately reproduces unpublished experimental work. Polyglutamine has been studied by FCS and simulation [116, 506, 508, 644]. Ash1 has been studied by SAXS and NMR (see 6. Polyglycine has been studied by FCS (see 5. sfAfp

[34]Gibbs et al. Manuscript*in press*

[35]More recently, unpublished smFRET studies have confirmed this result

has been studied by SAXS under oxidizing conditions where the internal disulphide bonds are broken [193]. All amino acid sequences can be found in appendix D.

The current iteration of GCF is able to accurately reproduce the global dimensions of a wide range of IDPs with surprisingly good accuracy. The three least accurate are highlighted with black arrows. For both DDX4N1 and NICD, experimental biophysical characterization of these IDPs has not yet been performed. For Ash1, we have *extensive* SAXS, NMR, and simulation data to strongly suggest that PIMMS is over-estimating the global dimensions of this sequence. However, by and large PIMMS + GCF is able to accurately reproduce highly sequence-specific global dimensions.

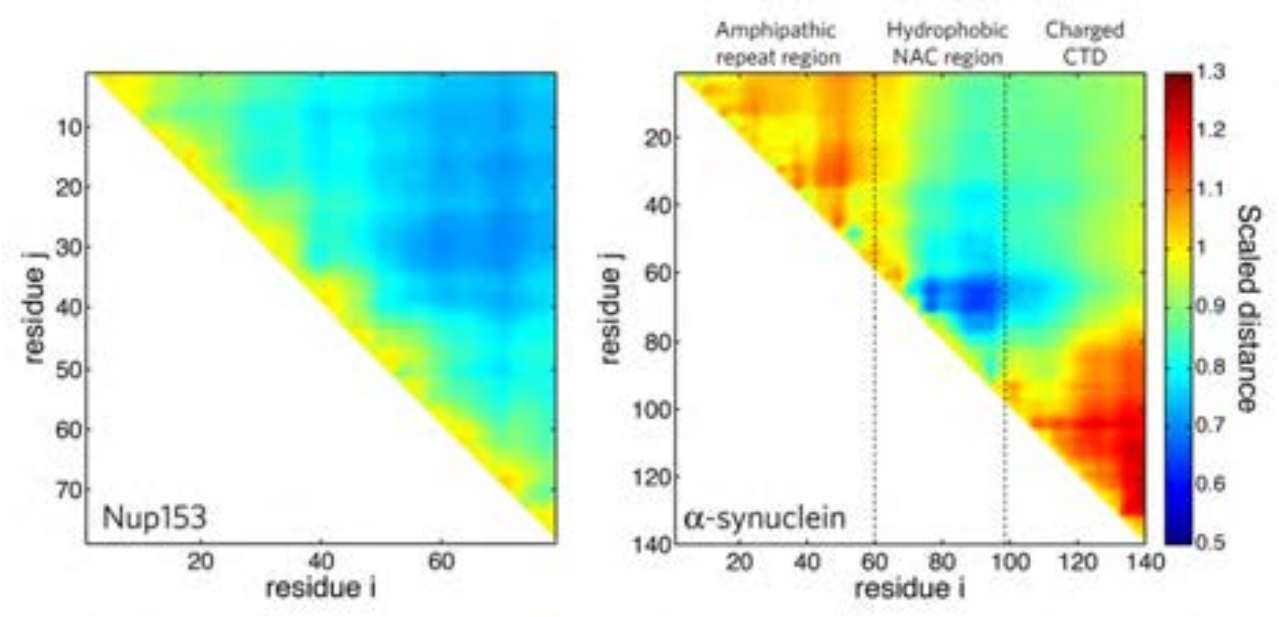

Figure 14.12: Scaling maps for Nup153 and α-synuclein show local and long-range conformational behaviour consistent with expectations based on previous work.

Are these results being obtained for the right reasons? This is a challenging question to answer. To address this, we analysed the distance map associated with two IDPs we expect to show entirely distinct local conformational behaviour, Nup153 and α-synuclein. Nup153

is part of the nuclear pore complex and is a relatively low complexity sequence, evenly peppered with phenylalanine residues. This sequences is believed to behave as a roughly uniformly expanded IDP, with several 'sticky patches' distributed across the sequence that can mediate weak multivalent interactions with partner proteins [378, 386]. In contrast, α-synuclein is known to contain three distinct regions; an amphipathic N-terminal repeat region, a hydrophobic central NAC region, and a negatively charged C-terminal region. The C-terminal region is believed to be relatively expanded under normal physiological pH [372]. As a result, α-synuclein is expected to show relatively sequence-specific conformational behaviour, while the local conformational behaviour of Nup153 is expected to be relatively uniform, perhaps with multiple distinct sites that engage in intermolecular interaction (again, a facsimile of Seminov and Rubinstein's 'stickers on a chain', see chapter 13 [532]). To compare these two sequences we generated scaling maps (as described in chapters 6 and 9), the results of which are shown in fig. 14.12.

Scaling map analysis produce two-dimensional scaling behaviour entirely consistent with our expectations. Nup153 is largely uniform, with multiple local minima in distance corresponding to the location of phenylalanine residues. Interestingly, not all phenylalanine residues are created equal, with the local sequence context playing an apparent modulatory role. For α-synuclein, the PIMMS simulations effectively delineate between these three regions, demonstrating they have distinct conformational properties consistent with our own unpublished all-atom simulations and various NMR studies. Arguably, the most important outcome from this is not that we obtain results that are consistent with our expectations, but that PIMMS simulations are able to uncover complex conformational behaviour with well defined sequence specificity.

It is important to not overstate these results. PIMMS remains a lattice based simulation engine where interactions are uniformly isotropic. Proteins exist in continuous space, and sidechains show well defined directionality. We are not arguing that PIMMS is correctly capturing the fine-grain detail of protein conformational behaviour. Instead, we propose PIMMS can act as a numerical version of Muthukumar's theory, abstracting the challenges of dealing with complex, n-ary systems and allowing emergent behaviour to appear organically.

14.3.7 Final Comments

PIMMS and GCF remain a work on progress, however they represent promising directions to ask an entire class of questions that are currently beyond the scope of most simulation engines. There are three major places we think PIMMS will provide novel insight.

Firstly, it provides a high-throughout tool to screen disordered regions on a proteome-scale to correlate primary sequence with conformational behaviour. Naturally this 'conformational behaviour' is a coarse-grained view of the true protein's behaviour, but nevertheless it will allow us to functionally annotate entire proteomes with biophysical insight. This in turn allows for a range of sequence-to-ensemble-to-function questions to be asked, as well as a set of questions pertaining to functional selection in proteins. More interesting, where we get things wrong implies the existence of strongly directional-dependent interactions or local structure.

Secondly, it allows us to ask a series of general questions regarding limiting models in phase separation. The 'stickers on a chain' model that has been discussed so frequently is still in principle a mean-field description [532]. Using PIMMS we can perform systematic investigations into a wide range of properties associated with 'stickers on a chain'. These investigations

will allow us to construct new analytical theory for the description and classification of phase separation in heteropolymers.

Finally, PIMMS allows us to probe the mesoscopic organization within a droplet by running simulations at the droplet concentration and examining the types of local structure observed. We suspect that the droplet interior is a complex mesh-work with large gaps separated by locally protein-dense regions undergoing extensive transient interactions. RNA may play a role in bridging local regions, or allowing droplets to become even more dilute. In summary, the use cases are plentiful, and we look forward to exploring a wide range of questions in biophysics and polymer chemistry using a robust, reliable, and efficient framework.

Chapter 15

Future directions II: Biological phase separation

The preceding chapters have introduced several ideas and discoveries pertaining to biological phase separation and condensate formation. Given the fact that the field is in a nascent stage, their are a plethora of possible questions to ask relating to the basic biophysics and cellular role of these processes. Many of the ideas that are discussed in this final chapter were introduced in the introductory chapter 3, so in considering the open questions we will be brief. Suffice to say, the questions posed here and in chapter 3 represent a small subset of the possible questions one could ask.

15.1 Sub-Diffraction Sized Biological Condensates

We hypothesize that small (50-200 mer) biological condensates may be associated with a wide range of processes. Can such a small assembly be 'liquid like'? We suggest that while lacking the long-range disorder of a true liquid, these finite-sized liquid phases could share many of

the features (a surface tension, rapid internal re-arrangement). It is also reasonable to ask if we are simply renaming macromolecular complexes as biomolecular condensates; this is certainly a possibility. As mentioned previously, we suggest that for these small, dynamic assemblies the term quinary assemblies becomes useful (see chapter 1), as it explicitly defines an assembly in which there is no fixed stoichiometry between the components [607, 622]. This is in direct contrast to 'conventional' macromolecular assemblies in which well defined stoichiometry exists between the various components (e.g. the F_1F_0-ATPase) [2]. Again, this is not currently the convention used, but we believe provides a useful framework to distinguish between different types of macromolecular assemblies.

Regardless, we believe that novel super-resolution techniques will allow the field to identify and describe the state and dynamics associated with these complexes, and determine if biology is truly using phase separation as a free-lunch for forming complex molecular assemblies *in vivo*. We believe the methods developed by the Cisse lab provide the spatial and temporal resolution to probe these nanoscopic dynamics, and are excited to follow as new insights emerge [100, 107, 412].

15.2 Spatial and Temporal Encoding

The ability of phase separation to drive spatial organization should be entirely obvious the dense phase(s) concentrates various components into a spatially distinct region. In addition we speculate that phase separation also provides a mechanism to allow control over temporal organization. The formation of condensates in response to some signal (e.g. heat stress) could facilitate the sequestration of a wide set of components into a liquid or gel-like

condensate. Once formed, those condensates could provide a kinetically stable local environment that traps the various components in place until the condensate is disassembled. From the perspective of the components absorbed by such a condensate, they would effectively 'jump' in time. Moreover, by forming condensates that undergo phase separation followed by gelation, this process could drive an irreversible sequestration that is only reversed via some active process. Such a mechanism provides an energy-independent, auto-responsive, and pseudo-irreversible mechanism for the sequestration of cellular components, the exact set of criterion the cell may want for a stress response. However, beyond the stress response, one could imagine various possible scenarios where such a change in cellular status could have a key functional role (differentiation, the propagation of cellular state, immunological sensitization *etc.*). For example, small, diffraction limited assemblies that form after some nucleation-limited process are at least a plausible explanation for recent work from Chakrabortee *et al.* [91]. The relationship between these assemblies and more conventional prion-based propagation remains an open and important question [283, 515].

15.3 Mixing, Specificity, and Local Organization

Many of the naturally occurring membrane-less organelles examined so far contain a large number of distinct components. Are these condensates homogenously mixed, or is a complex internal substructure present? We suspect the latter, but taken to its extreme it is somewhat unclear why well defined microdomains of distinct composition do not form inside macroscopic droplets. One possible explanation is they do, that these microdomains are sub-diffraction limited, and a word of nanometer inhomogeneities awaits our discovery. The

other is that they don't, and some kind of active process helps pull the system from equilibrium and away from a coarsened poorly-mixed state. This is also entirely possible, given varying lines of evidence that suggest internal dynamics are dependent on ATP, suggesting either that ATP is being directly used, or that ATP itself aids in maintaining local chain dynamics [261, 444, 639].

A final distinct but related question pertains to the molecular determinants of condensate accessibility and specificity. Nott *et al.* explored the molecular determinants of specificity in liquid droplets formed by the N-terminal domain of the DDX4 protein [420]. In this work, they found a correlation between amino acid composition and preferential partitioning. These results suggests that proteins could be 'addressed' to condensates simply by having a prerequisite disordered domain with the appropriate amino acid composition. Although an attractive hypothesis, this fails to offer a mechanism that allows any given low complexity sequence to distinguish between different droplets of similar composition, suggesting more work is required here.

15.4 Interplay of Folded and Disordered Domains

When early and ground-breaking work by Kato, Han, and Kwon was first published, it set the stage for intense focus on the relevant properties associated with low complexity domains that allows them to form condensates [217,282,312,313]. This line of investigation has proved immensely fruitful. However, in the last year so it has become clear that the folded domains associated with these low complexity sequences - far from 'just going along for the ride' - can play a key and sometimes sufficient role for condensate formation [463, 483][36]. Many

[36]Franzmann *et al.* (unpublished)

questions remain with respect to the interplay between folded and unfolded domains. Do folded domains drive phase separation? Do they modulate the internal structure of condensates? Are there specific interfaces that drive interaction, or are folded domains uniformly sticky? To what extent do folded domains and disordered regions function cooperatively, or are they largely independent entities? How we approach and answer these questions will define progress in the field, and a better integration of the results determined *in vitro* into their true biological contexts.

Appendix A

PIMMS keywords

Definition of keywords used for PIMMS simulations. Not included are the set of keywords that define the move-frequencies (one keyword per move-type).

Keyword	Meaning
DIMENSIONS	Size of the simulation box (in lattice units). 2D or 3D (defines if the simulation is a 2D or 3D simulation)
CHAIN	One of the few multi-component keywords in PIMMS and the only keyword that can appear multiple times, the CHAIN keyword defines a specific polymer chain and the number of that chain that will exist in the simulation. The format should be `CHAIN : N [CHAIN IDENTITY]` Where `N` defines the number of the chain and `CHAIN IDENTITY` defines the polymer sequence in one-letter alphabet code. As an example `CHAIN : 20 QQQQQQQQQQ` would define the system as having give 20 Q_{10} polymers.
TEMPERATURE	Simulation temperature in arbitrary units. The meaning of temperature will depend on on the strength of interactions used. All interaction energies are defined as integers to aid in numerical precision, so the current iterations of PIMMS parameterization uses T = 50.
N_STEPS	Total number of Monte Carlo steps to perform
PARAMETER_FILE	Location of parameter file to be used for the simulation

Table A.1: Summary of PIMMS keywords used for simulation configuration (1/2)

Keyword	Meaning
EQUILIBRATION	Number of steps to discard as equilibration
NON_INTERACTING	If set to true, all non-bonded interactions are switched off
ANGLES_OFF	If set to true, all torsional potential terms are switched off
PRINT_FREQ	Frequency with which general status information is printed to STDOUT
XTC_FREQ	Frequency with which the lattice configuration is written to the output trajectory file
EN_FREQ	Frequency which with system energy is written to an output file
SEED	Random-seed to be used (if a seed is not provided then the system generates a random seed). This allows fully reproducible simulations to be run.
ENERGY_CHECK	Frequency with which the current energy is compared to the *de novo* calculated global energy
ANALYSIS_FREQ	Frequency with which analysis is performed. For specific types of analysis this can be overwritten by an appropriate
TSMMC_JUMP_TEMP	Temperature maximum temperature TSMMC moves will reach
TSMMC_STEP_MULTIPLIER	Number of moves per temperature step on the TSMMC auxiliary chain
TSMMC_INTERPOLATION_MODE	Functional form of the temperature transition (currently linear is the only option)
TSMMC_NUMBER_OF_POINTS	Number of temperature steps between production temperature and jump temperature
TSMMC_FIXED_OFFSET	Set jump temperature as a fixed temperature offset from the main system temperature

Table A.2: Summary of PIMMS keywords used for simulation configuration (1/2)

Keyword	Meaning
ANA_*	There are a variety of ANA_* keywords that simply define the frequency with which certain types of analysis are performed, which we have not included here in the interest of brevity
QUENCH_RUN	Boolean to define if the simulation is to be run as a temperature quench or not
QUENCH_FREQ	If a quench run is performed, this sets the frequency with which the system temperature is updated
QUENCH_STEPSIZE	If a quench run is performed, this defines the change in temperature that occurs upon each update
QUENCH_START	If a quench run is performed, this defines the starting temperature
QUENCH_END	If a quench run is performed, this defines the ending temperature (this will be the production temperature and will override TEMPERATURE
QUENCH_AS_EQUILIBRATION	Boolean to set if the quench period should be treated as the equilibration period
CRANKSHAFT_SUBSTEPS	Number of individual crankshaft operations performed on each chain during a crankshaft move
CRANKSHAFT_MODE	How does the number of individual moves scale with chain length (can be none, linear, squared, cubed)

Table A.3: Summary of PIMMS keywords used for simulation configuration (3/3)

Appendix B

TSMMC: Derivation

The following represents a derivation of the acceptance correction factor used in for the temperature sweep Metropolis Monte Carlo (TSMMC). Please note that at the time of writing this is correct and maintains detailed balance. However, a more simple final form may be possible.

For TS-MMC to be a useful move in Monte Carlo simulations of macromolecules, it must be implemented in a manner that maintains micro-reversibility - i.e.

$$\pi_i p_{ij} = \pi_j p_{ji} \tag{B.1}$$

Here π is the probability of state i and p_{ij} is the probability of moving from state i to state j. p_{ij} is further defined as

$$p_{ij} = q_{ij}\alpha_{ij} \tag{B.2}$$

Where q_{ij} is the transition-matrix probability (the probability of selecting a move which will facilitate the i to j transition), while α_{ij} is the move's acceptance criterion (the probability of the move being accepted). In this work we use the Metropolis-Hasting acceptance, but in principle the following approach could be extended to use an alternative acceptance criterion. $\alpha_{i,j}$ is defined as

$$\alpha_{ij} = \frac{s_{ij}}{\left(1 + \frac{\pi_i}{\pi_j}\frac{q_{ij}}{q_{ji}}\right)} \tag{B.3}$$

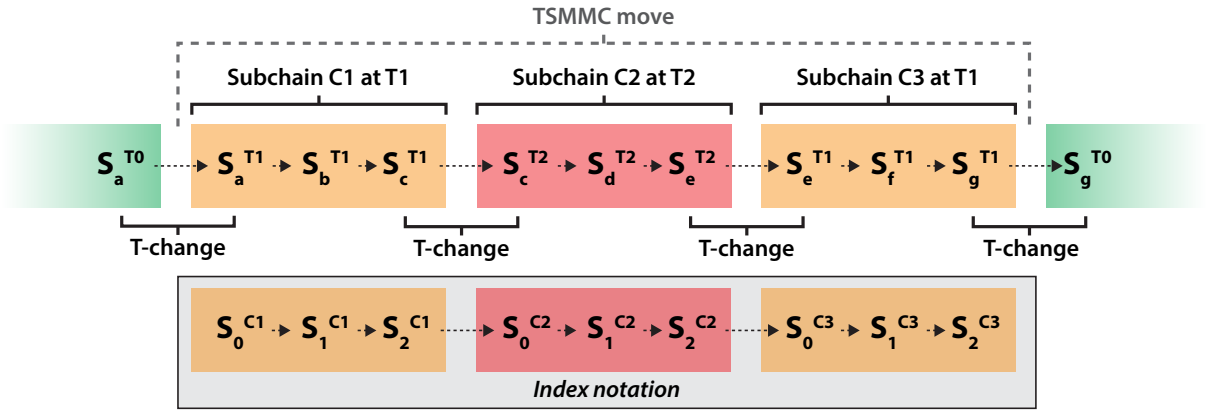

Figure B.1: A simple example of a TS-MMC move. The move facilitates the transition of state a to state g through T1 $\rightarrow$ T2 $\rightarrow$ T1 (above). The second chain (below) is the same TS-MMC chain written in index notation. This is a less clear notation, but allows for the transition probabilities to be written as a double sum (see eq. B.7), demonstrating the generality of the ideas. For the majority of the derivation we try to use the more convenient state-based notation, but switch to the index notation where necessary.

Where s_{ij} is

$$s_{ij} = \min\left(\left(1 + \frac{\pi_i}{\pi_j}\frac{q_{ij}}{q_{ji}}\right), \left(1 + \frac{\pi_j}{\pi_i}\frac{q_{ji}}{q_{ij}}\right)\right) \tag{B.4}$$

α_{ij} can therefore be written as

$$\begin{aligned}\alpha_{ij} &= \frac{\min\left(\left(1 + \frac{\pi_i}{\pi_j}\frac{q_{ij}}{q_{ji}}\right), \left(1 + \frac{\pi_j}{\pi_i}\frac{q_{ji}}{q_{ij}}\right)\right)}{\left(1 + \frac{\pi_i}{\pi_j}\frac{q_{ij}}{q_{ji}}\right)} \\ &= \min\left(\frac{\pi_j q_{ji}}{\pi_i q_{ij}}, 1\right)\end{aligned} \tag{B.5}$$

In the context of the TS-MMC we wish to transition from a state before and after the move. Let us use a specific example to define the auxiliary Markov chains, where we wish to transition between state a and state g (see fig. B.1).

For a given temperature ($T0$) The probability of moving from state a to state g can be re-written as

$$p_{a(T0)g(T0)} = q_{a(T0)g(T0)}\alpha_{a(T0)g(T0)} \tag{B.6}$$

Our objective is to define q and α in the context of the TS-MMC move such that micro-reversibility is maintained. $q_{a(T0)g(T0)}$ is the probability of selecting the series of moves which allow for the transition from state a to g. Note that these moves happen at different temperatures, but the transition between temperatures are accepted with a probability of 1.0 after a defined number of steps, so there is no need to consider a 'temperature change' move. For

mathematical convenience, we will represent states as indices (rather than letters), where for our example state $a = 0$ and $g = 2$. Similarly, we will represent temperatures as separate chain indices Ci. For a comparison of the mathematical index notation used vs. the more intuitive state and temperature notation see fig. B.1.

As a result, $q_{a(T0)g(T0)}$ can be written as follows

$$\begin{aligned} q_{a(T0)g(T0)} &= q_{0(C1)2(C3)} \\ &= \prod_{j=1}^{N_c} \prod_{i=1}^{N_i} p_{[i-1](Cj),i(Cj)} \end{aligned} \tag{B.7}$$

Here N_c is the number of sub-chains (in our example from fig. B.1 this would be 3) and N_i is the number of individual MMC moves perform within each chain. From a notation perspective it is convenient to have the same number of MMC moves within each subchain, but providing the same number of moves are performed for the equivalent temperature during the heating and cooling halves of the TS-MMC procedure there is no formal requirement for all subchains to perform the same number of moves.

We can next use B.2 and B.5 to re-write equation B.7,

$$
\begin{aligned}
q_{a(T0)g(T0)} = & \\
& = q_{0(C1)2(C3)} \\
& = \prod_{j=1}^{N_c} \prod_{i=1}^{N_i} p_{[i-1](Cj),i(Cj)} \\
& = \prod_{j=1}^{N_c} \prod_{i=1}^{N_i} q_{[i-1](Cj),i(Cj)} \ \alpha_{[i-1](Cj),i(Cj)} \\
& = \prod_{j=1}^{N_c} \prod_{i=1}^{N_i} q_{[i-1](Cj),i(Cj)} \ \min\left(\frac{\pi_{i(Cj)}}{\pi_{[i-1](Cj)}} \frac{q_{i(Cj),[i-1](Cj)}}{q_{[i-1](Cj),i(Cj)}}, 1 \right)
\end{aligned} \tag{B.8}
$$

In summary, the probability of the series of moves between states a and g (i.e. $0(C1)\ 2(C3)$) and is the product of the probability of the sub-moves in each of sub-chains (assuming each move therein respects micro-reversibility). Recall this is determined by the probability of the transition matrix associated with these moves, but not the likelihood of them being accepted, which is governed by α.

We now need a way to define α, which (recall equation B.3) can be written as

$$
\alpha_{a(T0)g(T0)} = \frac{s_{a(T0)g(T0)}}{1 + \dfrac{\pi_{a(T0)}}{\pi_{g(T0)}} \dfrac{q_{a(T0)g(T0)}}{q_{g(T0)a(T0)}}} \tag{B.9}
$$

Given can that $q_{a(T0)g(T0)}$ was defined in equation B.8 we can substantially simplify the $\frac{q_{a(T0)g(T0)}}{q_{g(T0)a(T0)}}$ component of equation B.9 $\left(\frac{q_{0(C1),2(C3)}}{q_{2(C3),0(C1)}} \text{ in index notation}\right)$ by canceling like terms associated with the summations in the denominator and numerator - i.e.

$$\frac{a \times \min\left(\frac{bx}{ay}, 1\right)}{b \times \min\left(\frac{ay}{bx}, 1\right)} = \frac{x}{y} \tag{B.10}$$

Where a and b are replaced by transition probabilities (q) while x and y are replaced by state probabilities (π). Practically, this is describing the ratio of the transition probabilities associated with traversing the TS-MMC path in the forward ($q_{0(C1),2(C3)}$) and backwards ($q_{2(C3),0(C1)}$) directions.

When all terms are canceled we are left with,

$$\frac{q_{a(T0)g(T0)}}{q_{g(T0)a(T0)}} = \frac{\pi_{g(T1)}}{\pi_{a(T1)}} \tag{B.11}$$

Which can be substituted into equation B.9 to give

$$\alpha_{a(T0)g(T0)} = \frac{s_{a(T0)g(T0)}}{1 + \frac{\pi_{a(T0)}}{\pi_{g(T0)}}\frac{\pi_{g(T1)}}{\pi_{a(T1)}}} \tag{B.12}$$

Finally we must define s_{ij}. Using eq. B.4 we can write,

$$s_{a(T0)g(T0)} = \min\left(\left(1 + \frac{\pi_{a(T0)}q_{a(T0)g(T0)}}{\pi_{g(T0)}q_{g(T0)a(T0)}}\right), \left(1 + \frac{\pi_{g(T0)}q_{g(T0)a(T0)}}{\pi_{a(T0)}q_{a(T0)g(T0)}}\right)\right) \tag{B.13}$$

This can be re-written as

$$s_{a(T0)b(T1)} = \min\left(\left(1 + \frac{\pi_{a(T0)}\pi_{g(T1)}}{\pi_{g(T0)}\pi_{a(T1)}}\right), \left(1 + \frac{\pi_{g(T0)}\pi_{a(T1)}}{\pi_{a(T0)}\pi_{g(T1)}}\right)\right) \tag{B.14}$$

Combining equation B.14 with equation B.12 we are left with

$$\alpha_{a(T0)g(T0)} = \min\left(\frac{\pi_{g(T0)}\pi_{a(T1)}}{\pi_{a(T0)}\pi_{g(T1)}}, 1\right) \tag{B.15}$$

This acceptance term acts as a correction to the full TS-MMC move. Recall that equation B.6 describes the probability of the TS-MMC move. The transition probability component of this ($q_{a(T0)g(T0)}$) is inherently captured by the fact that the moves within the TS-MMC subchains individually fulfil micro-reversibility, and because the subchain structure is symmetrical (in terms of the temperature changes) in the forwards and backwards direction. Consequently, the acceptance component ($\alpha_{a(T0)g(T0)}$) can be thought of as a correction factor to the full TS-MMC move after the move has complete. This derivation and justification is analogous to previously derivations from Gelb and Mittal *et al* [196, 394].

$$P(\text{accept}) \propto \frac{\exp\left(-\frac{E_g}{T0}\right) \times \exp\left(-\frac{E_a}{T1}\right)}{\exp\left(-\frac{E_a}{T0}\right) \times \exp\left(-\frac{E_g}{T1}\right))} \tag{B.16}$$

In conclusion, the accept/reject criterion for the TS-MMC move relies on a correction from the penultimate temperature. The TS-MMC move is implemented as a series of stand-alone Markov chains at varying temperatures with no need for accept/reject the actual temperature

changes. As a result the acceptance probability of the move is strongly linked to the extent of temperature change and the number of MMC moves performed within each chain. While the move is inherently more expensive than simple MMC moves that perturb a single degree of freedom, these two parameters can be tuned for a given system to maximize acceptance of the TS-MMC move.

Appendix C

The Amino Acids

Proteins are polypeptides - polymers of amino acids (the monomers) connected by peptide bonds. The peptide (amide) bond, highlighted in figure C.1 connects the COOH group of one amino acid to the NH2 group of the next. By convention, amino acid sequences are written in the N-to-C direction, i.e. -(-NH2-[Cα]-COOH-NH2-[Cα]-COOH-)-. Proteins can range in length from 30 - 40 residues (WW domain) to 35,000 residues (Titin). Each amino acid contains a central α-carbon flanked on one side by an amino (NH2) group and on the other by a carboxyl (COOH) group.

From the the central α-carbon atom extends the 'R' group (also referred to as the sidechain). There are twenty naturally occurring amino acids, each possessing a different sidechain with different physiocochemical properties. These chemical diversity provided by the sidechains includes differences in hydrophobicity, charge, size, aromaticity and hydrogen bonding potential. As a result, the linear combination of twenty chemically distinct monomers without constraints on the order those monomers appear in or the length of the resulting polymer provides nature with the opportunity to construct complex and specific heteropolymers. Importantly, the cellular machinery responsible for transcribing (converting a DNA sequence

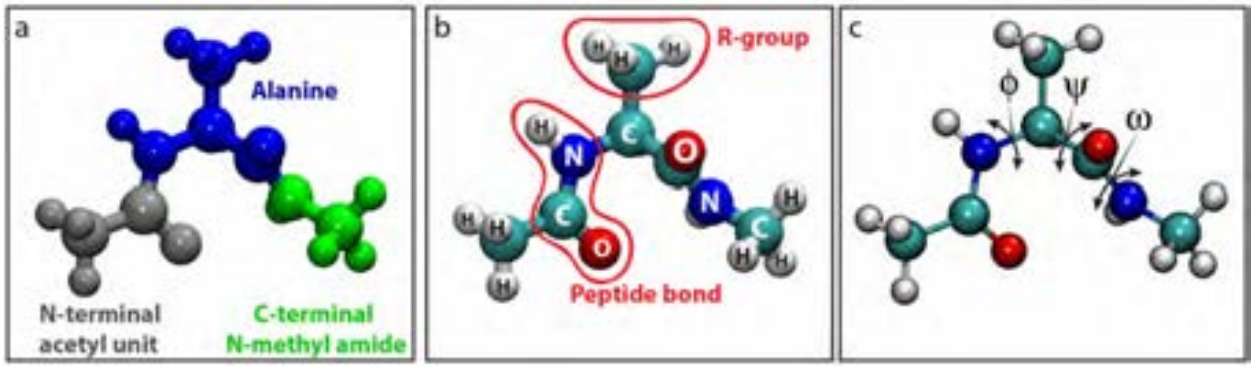

Figure C.1: The molecular structure of an amino acid. (a) The alanine dipeptide is frequently used as the hydrogen atom of protein chemistry. It consists of an alanine residue (blue) flanked by a capping N-terminal acetyle unit and an C-terminal N-methyl amide. (b) The peptide bond and sidechain (or R-group) represent two of the defining features of an amino acid. (c) The three key backbone bonds (ϕ, ψ, and ω are shown.

into an mRNA molecule) and translating (converting that mRNA molecule into a polypeptide) acts with an incredible fidelity, such that these complex heteropolymers are reproducibly and robustly synthesized with perfect precision.

In addition to the chemical diversity provided by the sidechains, the peptide backbone can engage in chemical interactions with sidechains, solvent, and solutes. There there is a significant degree of electron delocalization from the lone pair on the peptide bond nitrogen, causing the carbon atom associated with the peptide bond to display partial SP2 hybridization. One consequence of this is that the peptide bond shows a strong planar geometry, typically existing in either a *cis* or a *trans* conformation. Another consequence is that the peptide bond displays some ability to engage in pi-based interactions, as well as hydrogen

bonding. We will return to the potential importance of this pi-based interaction in later chapters, as well as a general discussion on the types of interactions the backbone can participate in.

There are twenty distinct amino acids, with their chemical structures shown in figure C.2. In the interest of completeness, we will provide a brief description of each of the amino acids below, divided into five groups chemically distinct groups.

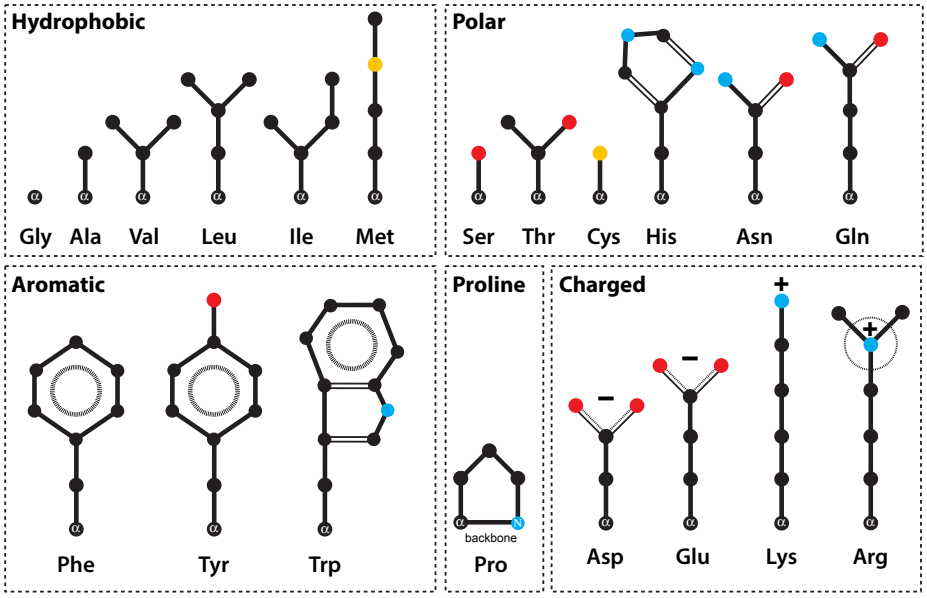

Figure C.2: The chemical structure of all twenty amino acids. Black denotes carbon, blue nitrogen, red oxygen, and yellow sulphur. Double bond characteristic is shown in a dashed line. Hydrogen atoms are not shown. The backbone cabon atom (Cα) is highlighted using a single α character. Proline is technically an *imino* acid, with the backbone carbon (Cα) and backbone nitrogen (N) contributing to the structure of the sidechain. Note that bond angles here are not realistic.

Hydrophobic Amino Acids

Alanine (Ala, A) has a small, somewhat hydrophobic sidechain and is one of the most abundant amino acids in most eukaryotic proteomes [402]. It has a strong propensity to form α-helices and is known to aggregate when found in polyalanine stretches [37]. **Isoleucine** (Ile, I). **Leucine** (Val, V) and **Leucine** (Leu, L) are all relatively large hydrophobic residue that are often associated with the formation of early hydrophobic contacts in protein folding (I/L/V clusters) [11, 192, 281, 371, 655]. **Methionine** (Met, M) is also a relatively large hydrophobic residue, sometimes considered less hydrophobic than I/L/V, and is typically observed at a much lower frequency than the other hydrophobic residues.

Aromatic Amino Acids

Phenylalanine (Phe, F) has an benzene aromatic side chain, making it highly hydrophobic and allowing it to engage in pi-pi and cation-pi interactions [150, 190]. **Tyrosine** (Tyr, Y) has a phenol aromatic ring, making it similar to phenylalanine in terms of its ability to engage in some pi-pi and cation-pi based interactions, as well as hydrogen bonding via the hydroxyl group. This combination of a pi-system and hydrogen bonding potential makes it a versatile, if relatively infrequent amino acid. **Tryptophan** (Trp, W) has a large doubly-ringed planar aromatic residue. It is traditionally considered to be the most hydrophobic of the three, is often involved in molecular recognition, and can in principle engage in pi-pi interactions, although is one of the least frequently observed amino acids [606].

Polar Amino Acids

Glycine (Gly, G) has no sidechain and is the only amino acid without a chiral center. The lack of steric interference from sidechains means the backbone amide group is significantly more accessible in glycine than in other amino acids, and typically glycine can engender increased flexibility to a polypeptide, leading to it capping and/or breaking α-helicies [17]. **Histadine** (His, H) has a pKa of 6.0, meaning that while at the standard cellular pH histadine is neutral, a small reduction in pH can lead to a substantial fraction of histadine residues being protonated. This makes it a good residue for cellular pH sensing , and for enzyme catalysis due to the ability to cycle charge states [197,215]. Additionally, it can engage in hydrogen bonding and pi-interactions due to the sp^2 characteristic of two of the carbon atoms. **Cysteine** (Cys, C) can form disulphide bonds, making it a good redox sensor. The thiol group generally does not act as a strong hydrogen bond donor, but it has a relatively strong propensity to form β-strands. **Asparagine** (Asn, N) contains a primary amide sidechain that can participate in hydrogen bonding and drives β-turn formation. Polyasparagine has been observed to undergo rapid aggregation *in vitro* [346]. **Glutamine** (Gln, Q) has a polar sidechain with an amide functional group that participates in hydrogen bonding. Gln can drives polypeptide chain collapse through sidechain-sidechain and sidechain-backbone interactions. While it does engage in favourable sidechain-water interactions, in the context of a polyglutamine tract Gln-Gln interactions (both sidechain-sidechain and sidechain-backbone) are preferred, leading to chain compaction. **Serine** (Ser, S) contains a short sidechain with a hydroxyl functional group that can participate in hydrogen bonding. Like serine, **Threonine** (Thr, T) contains a hydroxyl functional group an a methyl group, making it more hydrophobic than serine. The exact impact of threonine and serine on polypeptides is less well understood.

Proline

Proline (Pro, P) receives its own group, as it imparts a number of distinctive properties into proteins [518]. Proline is technically an imino acid. It is extremely soluble, and due to it's structural properties imparts significant inflexibility into the protein backbone, increasing the apparent persistence length [19, 643]. It is also a strong helix breaker, disrupting α-helices, and has a strong preference for the PPII helix and the coil state in general [17].

Charged Amino Acids

Aspartic acid (Asp, D) has a pKa of 3.71, meaning under normal cellular conditions it is negatively charged. **Glutamic Acid** (Glu, E) is similar with a pKa of 4.15, meaning it too is negatively charged under normal cellular conditions. Additionally, it shows an extremely strong preference for the formation of left-handed α-helices [112]. **Lysine** (Lys, K) has a positively charged amino acid with a pKa of 10.67. The sidechain contains a long hydrophobic methylene chain with a charge terminal amino group, giving it a bipartite, detergent-like chemical structure. **Arginine** (Arg, R) contains a positively charged sidechain with guanidinium functional group. The pKa of this sidechain is typically estimated to be ¿12, and it's neutralization is almost impossible. The central carbon in this guanidinium group has a significant SP2 characteristic, meaning it participates in significant pi based interactions as well as electrostatic interactions. This engenders a range of length-scales over where arginine can interact with other chemical moieties.

Appendix D

IDPs Used in PIMMS Simulations

RGG

MESNQSNNGG SGNAALNRGG RYVPPHLRGG DGGAAAAASA GGDDRRGGAG

GGGYRRGGGN SGGGGGGGYD RGYNDNRDDR DNRGGSGGYG RDRNYEDRGY

NGGGGGGGNR GYNNNRGGGG GGYNRQDRGD GGSSNFSRGG YNNRDEGSDN

RGSGRSYNND RRDNGGDG

DDX4

MGDEDWEAEI NPHMSSYVPI FEKDRYSGEN GDNFNRTPAS SSEMDDGPSR

RDHFMKSGFA SGRNFGNRDA GECNKRDNTS TMGGFGVGKS FGNRGFSNSR

FEDGDSSGFW RESSNDCEDN PTRNRGFSKR GGYRDGNNSE ASGPYRRGGR

GSFRGCRGGF GLGSPNNDLD PDECMQRTGG LFGSRRPVLS GTGNGDTSQS

RSGSGSERGG YKGLNEEVIT GSGKNSWKSE AEGGES

SYN

MDVFMKGLSK AKEGVVAAAE KTKQGVAEAA GKTKEGVLYV GSKTKEGVVH

GVATVAEKTK EQVTNVGGAV VTGVTAVAQK TVEGAGSIAA ATGFVKKDQL

GKNEEGAPQE GILEDMPVDP DNEAYEMPSE EGYQDYEPEA

CTD

FAGSGSNIYS PGNAYSPSSS NYSPNSPSYS PTSPSYSPSS PSYSPTSPCY

SPTSPSYSPT SPNYTPVTPS YSPTSPNYSA SPQ

Sic1

GSMTPSTPPR SRGTRYLAQP SGNTSSSALM QGQKTPQKPS QNLVPVTPST

TKSFKNAPLL APPNSNMGMT SPFNGLTSPQ RSPFPKSSVK RT

PAB1

YQQATAAAAA AAAGMPGQFM PPMFYGVMPP RGVPFNGPNP QQMNPMGGMP

KNGMPPQFRN GPVYGVPPQG GFPRNANDNN

Nup153

GCPSASPAFG ANQTPTFGQS QGASQPNPPG FSISSSTALF PTGSQPAPPT

GTVSSSSQPP VFGQQPSQSA FGSTTPNA

NICD

NASCVGGVLW QRRLRRLAEG ISEKTEAGSE EDRVRNEYEE SQWTGERDTQ

SSTVSTTEAE PYYRSLRDFS PQLPPTQEEV SYSRGFTGED EDMAFPGHLY

DEVERTYPPS GAWGPLYDEV QMGPWDLHWP EDTYQDPRGI YDQVAGDLDT

LEPDSLPFEL RGHLV

SV1

EKEKEKEKEK EKEKEKEKEK EKEKEKEKEK EKEKEKEKEK EKEKEKEKEK

EGFR

MERMHLPSPT DSNFYRALMD EEDMDDVVDA DEYLIPQQGF FSSPSTSRTP

LLSSLSATSN NSTVACIDRN GLQSCPIKED SFLQRYSSDP TGALTEDSID

DTFLPVPEYI NQSVPKRPAG SVQNPVYHNQ PLNPAPSRDP HYQDPHSTAV

GNPEYLNTVQ PTCVNSTFDS PAHWAQKGSH QISLDNPDYQ QDFFPKEAKP

NGIFKGSTAE NAEYLRVAPQ SSEFIGALEH HHHHH

polyQ

QQQQQQQQQQ QQQQQQQQQQ QQQQQQQQQQ QQQQQQ

SV30

EEEEEEEEEE EEEEEEEEEE EEEEEKKKKK KKKKKKKKKK KKKKKKKKKK

polyG

GGGGGGGGGG GGGGG

ASH1

GASASSSPSP STPTKSGKMR SRSSSPVRPK AYTPSPRSPN YHRFALDSPP

QSPRRSSNSS ITKKGSRRSS GSSPTRHTTR VCV

sfAFP

CKGADGAHGV NGCPGTAGAA GSVGGPGCDG GHGGNGGNGN PGCAGGVGGA

GGASGGTGVG GRGGKGGSGT PKGADGAPGAP

References

[1] Mark James Abraham, Teemu Murtola, Roland Schulz, Szilárd Páll, Jeremy C Smith, Berk Hess, and Erik Lindahl. GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX*, 1–2:19–25, 2015.

[2] Jan Pieter Abrahams, Andrew G W Leslie, René Lutter, and John E Walker. Structure at 2.8 a resolution of F1-ATPase from bovine heart mitochondria. *Nature*, 370(6491):621–628, 1994.

[3] P Ahlrichs, R Everaers, and B Dünweg. Screening of hydrodynamic interactions in semidilute polymer solutions: a computer simulation study. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.*, 64(4 Pt 1):040501, October 2001.

[4] Tural Aksel and Doug Barrick. Direct observation of parallel folding pathways revealed using a symmetric repeat protein system. *Biophys. J.*, 107(1):220–232, 1 July 2014.

[5] Bruce Alberts, Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts, and Peter Walter. *Molecular Biology of the Cell*. Garland Science, 2002.

[6] K Almdal, J Dyre, S Hvidt, and Ole Kramer. Towards a phenomenological definition of the term gel. *Polym. Gels Networks*, 1(1):5–17, 1993.

[7] Matthias Altmeyer, Kai J Neelsen, Federico Teloni, Irina Pozdnyakova, Stefania Pellegrino, Merete Grøfte, Maj-Britt Druedahl Rask, Werner Streicher, Stephanie Jungmichel, Michael Lund Nielsen, and Jiri Lukas. Liquid demixing of intrinsically disordered proteins is seeded by poly(ADP-ribose). *Nat. Commun.*, 6:8088, 19 August 2015.

[8] Jens S Andersen, Yun W Lam, Anthony K L Leung, Shao-En Ong, Carol E Lyon, Angus I Lamond, and Matthias Mann. Nucleolar proteome dynamics. *Nature*, 433(7021):77–83, 6 January 2005.

[9] C B Anfinsen. Principles that govern the folding of protein chains. *Science*, 181(4096):223–230, 20 July 1973.

[10] B Anil, S Sato, J H Cho, and D P Raleigh. Fine structure analysis of a protein folding transition state; distinguishing between hydrophobic stabilization and specific packing. *J. Mol. Biol.*, 354(3):693–705, 2005.

[11] M Arai, M Iwakura, C R Matthews, and O Bilsel. Microsecond subdomain folding in dihydrofolate reductase. *J. Mol. Biol.*, 410(2):329–342, 2011.

[12] M Arai, E Kondrashkina, C Kayatekin, C R Matthews, M Iwakura, and O Bilsel. Microsecond hydrophobic collapse in the folding of escherichia coli dihydrofolate reductase, an alpha/beta-type protein. *J. Mol. Biol.*, 368(1):219–229, 2007.

[13] Yukinobu Arata, Hiroaki Takagi, Yasushi Sako, and Hitoshi Sawa. Power law relationship between cell cycle duration and cell volume in the early embryonic development of caenorhabditis elegans. *Front. Physiol.*, 5:529, 2014.

[14] V L Arcus, S Vuilleumier, S M V Freund, M Bycroft, and A R Fersht. A comparison of the ph, urea, and temperature-denatured states of barnase by heteronuclear nmr - implications for the initiation of protein-folding. *J. Mol. Biol.*, 254(2):305–321, 24 November 1995.

[15] Timothy E Audas, Danielle E Audas, Mathieu D Jacob, J J David Ho, Mireille Khacho, Miling Wang, J Kishan Perera, Caroline Gardiner, Clay A Bennett, Trajen Head, Oleksandr N Kryvenko, Mercé Jorda, Sylvia Daunert, Arun Malhotra, Laura Trinkle-Mulcahy, Mark L Gonzalgo, and Stephen Lee. Adaptation to stressors by systemic protein amyloidogenesis. *Dev. Cell*, 39(2):155–168, 24 October 2016.

[16] William M Aumiller, Jr. and Christine D Keating. Phosphorylation-mediated RNA/peptide complex coacervation as a model for intracellular liquid organelles. *Nat. Chem.*, 8(2):129–137, February 2016.

[17] R Aurora and G D Rose. Helix capping. *Protein Sci.*, 7(1):21–38, January 1998.

[18] Matthew Auton and D Wayne Bolen. Additive transfer free energies of the peptide backbone unit that are independent of the model compound and the choice of concentration scale. *Biochemistry*, 43(5):1329–1342, 10 February 2004.

[19] Matthew Auton, Luis Marcelo F Holthauzen, and D Wayne Bolen. Anatomy of energetic changes accompanying urea-induced protein denaturation. *Proc. Natl. Acad. Sci. U. S. A.*, 104(39):15317–15322, 18 September 2007.

[20] Mikayel Aznauryan, Leonildo Delgado, Andrea Soranno, Daniel Nettels, Jie-Rong Huang, Alexander M Labhardt, Stephan Grzesiek, and Benjamin Schuler. Comprehensive structural and dynamical view of an unfolded protein from the combination of single-molecule FRET, NMR, and SAXS. *Proc. Natl. Acad. Sci. U. S. A.*, 113(37):E5389–98, 13 September 2016.

[21] M Madan Babu, Richard W Kriwacki, and Rohit V Pappu. Structural biology. versatility from protein disorder. *Science*, 337(6101):1460–1461, 21 September 2012.

[22] Marco Bacci, Andreas Vitalis, and Amedeo Caflisch. A molecular simulation protocol to avoid sampling redundancy and discover new states. *Biochim. Biophys. Acta*, 2 September 2014.

[23] Alaji Bah, Robert M Vernon, Zeba Siddiqui, Mickaël Krzeminski, Ranjith Muhandiram, Charlie Zhao, Nahum Sonenberg, Lewis E Kay, and Julie D Forman-Kay. Folding of an intrinsically disordered protein by phosphorylation as a regulatory switch. *Nature*, 519(7541):106–109, 5 March 2015.

[24] Xiao-Chen Bai, Greg McMullan, and Sjors H W Scheres. How cryo-EM is revolutionizing structural biology. *Trends Biochem. Sci.*, 40(1):49–57, January 2015.

[25] Y Bai, T R Sosnick, L Mayne, and S W Englander. Protein folding intermediates: native-state hydrogen exchange. *Science*, 269(5221):192–197, 14 July 1995.

[26] Andrew J Baldwin, Tuomas P J Knowles, Gian Gaetano Tartaglia, Anthony W Fitzpatrick, Glyn L Devlin, Sarah Lucy Shammas, Christopher A Waudby, Maria F Mossuto, Sarah Meehan, Sally L Gras, John Christodoulou, Spencer J Anthony-Cahill, Paul D Barker, Michele Vendruscolo, and Christopher M Dobson. Metastability of native proteins and the phenomenon of amyloid formation. *J. Am. Chem. Soc.*, 133(36):14160–14163, 14 September 2011.

[27] Salman F Banani, Hyun O Lee, Anthony A Hyman, and Michael K Rosen. Biomolecular condensates: organizers of cellular biochemistry. *Nat. Rev. Mol. Cell Biol.*, 22 February 2017.

[28] Salman F Banani, Allyson M Rice, William B Peeples, Yuan Lin, Saumya Jain, Roy Parker, and Michael K Rosen. Compositional control of Phase-Separated cellular bodies. *Cell*, 166(3):651–663, 28 July 2016.

[29] S Banjade and M K Rosen. Phase transitions of multivalent proteins can promote clustering of membrane receptors. *Elife*, 3, 2014.

[30] Sudeep Banjade, Qiong Wu, Anuradha Mittal, William B Peeples, Rohit V Pappu, and Michael K Rosen. Conserved interdomain linker promotes phase separation of the multivalent adaptor protein nck. *Proc. Natl. Acad. Sci. U. S. A.*, 112(47):E6426–35, 24 November 2015.

[31] A A Barker. Monte carlo calculations of the radial distribution functions for a Proton-Electron plasma. *Aust. J. Phys.*, 18(2):119–134, 1 April 1965.

[32] M Bastidas, E B Gibbs, D Sahu, and S A Showalter. A primer for carbon-detected NMR applications to intrinsically disordered proteins in solution. *Concepts Magn. Reson. Part A Bridg. Educ. Res.*, 44(1):54–66, 2015.

[33] Leslie Y Beh, Lucy J Colwell, and Nicole J Francis. A core subunit of polycomb repressive complex 1 is broadly conserved in function but not primary sequence. *Proc. Natl. Acad. Sci. U. S. A.*, 109(18):E1063–71, 1 May 2012.

[34] H J C Berendsen, J P M Postma, W F van Gunsteren, and J Hermans. Interaction models for water in relation to protein hydration. In Bernard Pullman, editor, *Intermolecular Forces*, The Jerusalem Symposia on Quantum Chemistry and Biochemistry, pages 331–342. Springer Netherlands, 1981.

[35] Jeremy M Berg, John L Tymoczko, and Lubert Stryer. *Biochemistry*. W H Freeman, 2002.

[36] H M Berman, J Westbrook, Z Feng, G Gilliland, T N Bhat, H Weissig, I N Shindyalov, and P E Bourne. The protein data bank. *Nucleic Acids Res.*, 28(1):235–242, 1 January 2000.

[37] Joseph P Bernacki and Regina M Murphy. Length-dependent aggregation of uninterrupted polyalanine peptides. *Biochemistry*, 50(43):9200–9211, 1 November 2011.

[38] P Bernado, M Blackledge, and J Sancho. Sequence-specific solvent accessibilities of protein residues in unfolded protein ensembles. *Biophys. J.*, 91(12):4536–4543, 2006.

[39] Pau Bernadó, Efstratios Mylonas, Maxim V Petoukhov, Martin Blackledge, and Dmitri I Svergun. Structural characterization of flexible proteins using small-angle x-ray scattering. *J. Am. Chem. Soc.*, 129(17):5656–5664, 2 May 2007.

[40] Rafael C Bernardi, Marcelo C R Melo, and Klaus Schulten. Enhanced sampling techniques in molecular dynamics simulations of biological systems. *Biochim. Biophys. Acta*, 1850(5):872–877, May 2015.

[41] Joel Berry, Stephanie C Weber, Nilesh Vaidya, Mikko Haataja, and Clifford P Brangwynne. RNA transcription modulates phase transition-driven nuclear body assembly. *Proceedings of the National Academy of Sciences*, 112(38):E5237–E5245, 22 September 2015.

[42] R B Best and G Hummer. Coordinate-dependent diffusion in protein folding. *Proc. Natl. Acad. Sci. U. S. A.*, 107(3):1088–1093, 2010.

[43] R B Best and J Mittal. Balance between a and beta structures in ab initio protein folding. *J. Phys. Chem. B*, 114(26):8790–8798, 2010.

[44] R B Best and J Mittal. Protein simulations with an optimized water model: Cooperative helix formation and Temperature-Induced unfolded state collapse. *J. Phys. Chem. B*, 114(46):14916–14923, 2010.

[45] Robert B Best, Gerhard Hummer, and William A Eaton. Native contacts determine protein folding mechanisms in atomistic simulations. *Proceedings of the National Academy of Sciences*, 110(44):17874–17879, 29 October 2013.

[46] Robert B Best, Wenwei Zheng, and Jeetain Mittal. Balanced Protein-Water interactions improve properties of disordered proteins and Non-Specific protein association. *J. Chem. Theory Comput.*, 10(11):5113–5124, 11 November 2014.

[47] Jan Bieschke, Martin Herbst, Thomas Wiglenda, Ralf P Friedrich, Annett Boeddrich, Franziska Schiele, Daniela Kleckers, Juan Miguel Lopez del Amo, Björn A Grüning, Qinwen Wang, Michael R Schmidt, Rudi Lurz, Roger Anwyl, Sigrid Schnoegl, Marcus Fändrich, Ronald F Frank, Bernd Reif, Stefan Günther, Dominic M Walsh, and Erich E Wanker. Small-molecule conversion of toxic oligomers to nontoxic β-sheet–rich amyloid fibrils. *Nat. Chem. Biol.*, 8(1):93–101, 20 November 2011.

[48] O Bilsel and C R Matthews. Molecular dimensions and their distributions in early folding intermediates. *Curr. Opin. Struct. Biol.*, 16(1):86–93, 2006.

[49] I M Blasutig, L A New, A Thanabalasuriar, T K Dayarathna, M Goudreault, S E Quaggin, S S Li, S Gruenheid, N Jones, and T Pawson. Phosphorylated YDXV motifs and nck SH2/SH3 adaptors act cooperatively to induce actin reorganization. *Mol. Cell. Biol.*, 28(6):2035–2046, 2008.

[50] Jesse D Bloom, Sy T Labthavikul, Christopher R Otey, and Frances H Arnold. Protein stability promotes evolvability. *Proc. Natl. Acad. Sci. U. S. A.*, 103(15):5869–5874, 11 April 2006.

[51] Steven Boeynaems, Elke Bogaert, Denes Kovacs, Albert Konijnenberg, Evy Timmerman, Alex Volkov, Mainak Guharoy, Mathias De Decker, Tom Jaspers, Veronica H Ryan, Abigail M Janke, Pieter Baatsen, Thomas Vercruysse, Regina-Maria Kolaitis, Dirk Daelemans, J Paul Taylor, Nancy Kedersha, Paul Anderson, Francis Impens, Frank Sobott, Joost Schymkowitz, Frederic Rousseau, Nicolas L Fawzi, Wim Robberecht, Philip Van Damme, Peter Tompa, and Ludo Van Den Bosch. Phase separation of c9orf72 dipeptide repeats perturbs stress granule dynamics. *Mol. Cell*, 65(6):1044–1055.e5, 16 March 2017.

[52] Elvan Boke, Martine Ruer, Martin Wühr, Margaret Coughlin, Regis Lemaitre, Steven P Gygi, Simon Alberti, David Drechsel, Anthony A Hyman, and Timothy J Mitchison. Amyloid-like Self-Assembly of a cellular compartment. *Cell*, 166(3):637–650, 28 July 2016.

[53] D Wayne Bolen and George D Rose. Structure and energetics of the hydrogen-bonded backbone in protein folding. *Annu. Rev. Biochem.*, 77:339–362, 2008.

[54] Heinerle L Booij and Hendrik G Bungenberg de Jong. Biocolloids and their interactions. In F Webe and L V Helbrunn, editors, *Protoplasmatologia: Handbuch der Protoplasmaforschung*. Springer, 1956.

[55] D R Booth, M Sunde, V Bellotti, C V Robinson, W L Hutchinson, P E Fraser, P N Hawkins, C M Dobson, S E Radford, C C Blake, and M B Pepys. Instability, unfolding and aggregation of human lysozyme variants underlying amyloid fibrillogenesis. *Nature*, 385(6619):787–793, 27 February 1997.

[56] Thomas C Boothby, Hugo Tapia, Alexandra H Brozena, Samantha Piszkiewicz, Austin E Smith, Ilaria Giovannini, Lorena Rebecchi, Gary J Pielak, Doug Koshland, and Bob Goldstein. Tardigrades use intrinsically disordered proteins to survive desiccation. *Mol. Cell*, 65(6):975–984.e5, 16 March 2017.

[57] Wade Borcherds, François-Xavier Theillet, Andrea Katzer, Ana Finzel, Katie M Mishall, Anne T Powell, Hongwei Wu, Wanda Manieri, Christoph Dieterich, Philipp Selenko, Alexander Loewer, and Gary W Daughdrill. Disorder and residual helicity alter p53-mdm2 binding affinity and signaling in cells. *Nat. Chem. Biol.*, 10(12):1000–1002, December 2014.

[58] M Borg, T Mittag, T Pawson, M Tyers, J D Forman-Kay, and H S Chan. Polyelectrostatic interactions of disordered ligands suggest a physical basis for ultrasensitivity. *Proc. Natl. Acad. Sci. U. S. A.*, 104(23):9650–9655, 2007.

[59] Alessandro Borgia, Wenwei Zheng, Karin Buholzer, Madeleine B Borgia, Anja Schüler, Hagen Hofmann, Andrea Soranno, Daniel Nettels, Klaus Gast, Alexander Grishaev, Robert B Best, and Benjamin Schuler. Consistent view of polypeptide chain expansion in chemical denaturants from multiple experimental methods. *J. Am. Chem. Soc.*, 138(36):11714–11726, 14 September 2016.

[60] B E Bowler. Thermodynamics of protein denatured states. *Mol. Biosyst.*, 3(2):88–99, 2007.

[61] Gregory R Bowman, Xuhui Huang, and Vijay S Pande. Using generalized ensemble simulations and markov state models to identify conformational states. *Methods*, 49(2):197–201, October 2009.

[62] Gregory R Bowman and Vijay S Pande. Protein folded states are kinetic hubs. *Proc. Natl. Acad. Sci. U. S. A.*, 107(24):10890–10895, 15 June 2010.

[63] H Boze, T Marlin, D Durand, J Perez, A Vernhet, F Canon, P Sarni-Manchado, V Cheynier, and B Cabane. Proline-rich salivary proteins have extended conformations. *Biophys. J.*, 99(2):656–665, 2010.

[64] Philip Bradley, Kira M S Misura, and David Baker. Toward high-resolution de novo structure prediction for small proteins. *Science*, 309(5742):1868–1871, 16 September 2005.

[65] Clifford P Brangwynne, Christian R Eckmann, David S Courson, Agata Rybarska, Carsten Hoege, Joebin Gharakhani, Frank Juelicher, and Anthony A Hyman. Germline P granules are liquid droplets that localize by controlled dissolution/condensation. *Science*, 324(5935):1729–1732, 26 June 2009.

[66] Clifford P Brangwynne, Timothy J Mitchison, and Anthony A Hyman. Active liquid-like behavior of nucleoli determines their size and shape in xenopus laevis oocytes. *Proc. Natl. Acad. Sci. U. S. A.*, 108(11):4334–4339, 2011.

[67] Clifford P Brangwynne, Peter Tompa, and Rohit V Pappu. Polymer physics of intracellular phase transitions. *Nat. Phys.*, 11(11):899–904, 3 November 2015.

[68] Amandaa K Brewer and André M Striegel. Characterizing the size, shape, and compactness of a polydisperse prolate ellipsoidal particle via quadruple-detector hydrodynamic chromatography. *Analyst*, 136(3):515–519, 7 February 2011.

[69] M L Broide, C R Berland, J Pande, O O Ogun, and G B Benedek. Binary-liquid phase separation of lens protein solutions. *Proc. Natl. Acad. Sci. U. S. A.*, 88(13):5660–5664, 1 July 1991.

[70] Celeste J Brown, Audra K Johnson, A Keith Dunker, and Gary W Daughdrill. Evolution and disorder. *Curr. Opin. Struct. Biol.*, 21(3):441–446, June 2011.

[71] Marco Brucale, Benjamin Schuler, and Bruno Samorì. Single-molecule studies of intrinsically disordered proteins. *Chem. Rev.*, 114(6):3281–3317, 26 March 2014.

[72] Bryan Marten, , Kyungsun Kim, ⊥ Christian Cortis, , Richard A. Friesner*, Robert B Murphy#, Murco N Ringnalda$, ∀ Doree Sitkoff∞, @, and Barry Honig*. New model for calculation of solvation free energies: Correction of Self-Consistent reaction field continuum dielectric theory for Short-Range Hydrogen-Bonding effects. *J. Phys. Chem.*, 100(28):11775–11788, 1996.

[73] J D Bryngelson, J N Onuchic, N D Socci, and P G Wolynes. Funnels, pathways, and the energy landscape of protein folding: a synthesis. *Proteins*, 21(3):167–195, March 1995.

[74] J D Bryngelson and P G Wolynes. Spin glasses and the statistical mechanics of protein folding. *Proc. Natl. Acad. Sci. U. S. A.*, 84(21):7524–7528, November 1987.

[75] J Ross Buchan. mRNP granules. assembly, function, and connections with disease. *RNA Biol.*, 11(8):1019–1030, 2014.

[76] J Ross Buchan, Regina-Maria Kolaitis, J Paul Taylor, and Roy Parker. Eukaryotic stress granules are cleared by autophagy and Cdc48/VCP function. *Cell*, 153(7):1461–1474, 20 June 2013.

[77] J Ross Buchan and Roy Parker. Eukaryotic stress granules: the ins and outs of translation. *Mol. Cell*, 36(6):932–941, 25 December 2009.

[78] Alexander K Buell, Céline Galvagnion, Ricardo Gaspar, Emma Sparr, Michele Vendruscolo, Tuomas P J Knowles, Sara Linse, and Christopher M Dobson. Solution conditions determine the relative importance of nucleation and growth processes in α-synuclein aggregation. *Proc. Natl. Acad. Sci. U. S. A.*, 111(21):7671–7676, 27 May 2014.

[79] Kathleen A Burke, Abigail M Janke, Christy L Rhine, and Nicolas L Fawzi. Residue-by-residue view of in vitro FUS granules that bind the c-terminal domain of RNA polymerase II. *Mol. Cell*, 60(2):231–241, 15 October 2015.

[80] Prasad V Burra, Lajos Kalmar, and Peter Tompa. Reduction in structural disorder and functional complexity in the thermal adaptation of prokaryotes. *PLoS One*, 5(8):e12069, 11 August 2010.

[81] Giovanni Bussi, Tatyana Zykova-Timan, and Michele Parrinello. Isothermal-isobaric molecular dynamics using stochastic velocity rescaling. *J. Chem. Phys.*, 130(7), 2009.

[82] Li-Heng Cai, Sergey Panyukov, and Michael Rubinstein. Mobility of nonsticky nanoparticles in polymer liquids. *Macromolecules*, 44(19):7853–7863, 11 October 2011.

[83] Xiao Dan Cai, Todd E Golde, and Steven G Younkin. Release of excess amyloid protein from a mutant amyloid protein precursor. *SCIENCE-NEW YORK THEN WASHINGTON-*, 259:514–514, 1993.

[84] C J Camacho and D Thirumalai. Modeling the role of disulfide bonds in protein folding: entropic barriers and pathways. *Proteins*, 22(1):27–40, 1995.

[85] C Camilloni, A De Simone, W F Vranken, and M Vendruscolo. Determination of secondary structure populations in disordered states of proteins using nuclear magnetic resonance chemical shifts. *Biochemistry*, 51(11):2224–2231, 2012.

[86] Andrew Campen, Ryan M Williams, Celeste J Brown, Jingwei Meng, Vladimir N Uversky, and A Keith Dunker. TOP-IDP-scale: A new amino acid scale measuring propensity for intrinsic disorder. *Protein Pept. Lett.*, 15(9):956–963, 2008.

[87] Deepak R Canchi and Angel E Garcia. Backbone and side-chain contributions in protein denaturation by urea. *Biophys. J.*, 100(6):1526–1533, 16 March 2011.

[88] Deepak R Canchi and Angel E García. Cosolvent effects on protein stability. *Annual Reviews of Physical Chemistry*, 64:273–293, 4 January 2013.

[89] M Carmo-Fonseca. The contribution of nuclear compartmentalization to gene regulation. *Cell*, 108(4):513–521, 2002.

[90] John Cavanagh, Wayne J Fairbrother, Iii Arthur G. Palmer, and Nicholas J Skelton. *Protein NMR Spectroscopy: Principles and Practice*. Academic Press, 28 November 1995.

[91] Sohini Chakrabortee, James S Byers, Sandra Jones, David M Garcia, Bhupinder Bhullar, Amelia Chang, Richard She, Laura Lee, Brayon Fremin, Susan Lindquist, and Daniel F Jarosz. Intrinsically disordered proteins drive emergence and inheritance of biological traits. *Cell*, 167(2):369–381.e12, 6 October 2016.

[92] C K Chan, Y Hu, S Takahashi, D L Rousseau, W A Eaton, and J Hofrichter. Submillisecond protein folding kinetics studied by ultrarapid mixing. *Proc. Natl. Acad. Sci. U. S. A.*, 94(5):1779–1784, 4 March 1997.

[93] H S Chan and K A Dill. Protein folding in the landscape perspective: Chevron plots and non-arrhenius kinetics. *Proteins-Structure Function and Bioinformatics*, 30(1):2–33, January 1998.

[94] Jessica Walton Chen, Pedro Romero, Vladimir N Uversky, and A Keith Dunker. Conservation of intrinsic disorder in protein domains and families: I. a database of conserved predicted disordered regions. *J. Proteome Res.*, 5(4):879–887, April 2006.

[95] Yan Chen, Joachim D Müller, Qiaoqiao Ruan, and Enrico Gratton. Molecular brightness characterization of EGFP in vivo by fluorescence fluctuation spectroscopy. *Biophys. J.*, 82(1 Pt 1):133–144, January 2002.

[96] Margaret S Cheung, Angel E García, and José N Onuchic. Protein folding mediated by solvation: water expulsion and formation of the hydrophobic core occur after the structural collapse. *Proc. Natl. Acad. Sci. U. S. A.*, 99(2):685–690, 22 January 2002.

[97] Alexander F Chin, Dmitri Toptygin, W Austin Elam, Travis P Schrank, and Vincent J Hilser. Phosphorylation increases persistence length and End-to-End distance of a segment of tau protein. *Biophys. J.*, 110(2):362–371, 19 January 2016.

[98] J H Cho and D P Raleigh. Mutational analysis demonstrates that specific electrostatic interactions can play a key role in the denatured state ensemble of proteins. *J. Mol. Biol.*, 353(1):174–185, 2005.

[99] Jae-Hyun Cho, Wenli Meng, Satoshi Sato, Eun Young Kim, Hermann Schindelin, and Daniel P Raleigh. Energetically significant networks of coupled interactions within an

unfolded protein. *Proc. Natl. Acad. Sci. U. S. A.*, 111(33):12079–12084, 19 August 2014.

[100] Won-Ki Cho, Namrata Jayanth, Brian P English, Takuma Inoue, J Owen Andrews, William Conway, Jonathan B Grimm, Jan-Hendrik Spille, Luke D Lavis, Timothée Lionnet, and Ibrahim I Cisse. RNA polymerase II cluster dynamics predict mRNA output in living cells. *Elife*, 5, 3 May 2016.

[101] John D Chodera and Frank Noé. Markov state models of biomolecular conformational dynamics. *Curr. Opin. Struct. Biol.*, 25:135–144, April 2014.

[102] John D Chodera, William C Swope, Jed W Pitera, Chaok Seok, and Ken A Dill. Use of the weighted histogram analysis method for the analysis of simulated and parallel tempering simulations. *J. Chem. Theory Comput.*, 3(1):26–41, 2007.

[103] Jeong-Mo Choi, Adrian W R Serohijos, Sean Murphy, Dennis Lucarelli, Leo L Lofranco, Andrew Feldman, and Eugene I Shakhnovich. Minimalistic predictor of protein binding energy: contribution of solvation factor to protein binding. *Biophys. J.*, 108(4):795–798, 17 February 2015.

[104] C Chothia and A M Lesk. The relation between the divergence of sequence and structure in proteins. *EMBO J.*, 5(4):823–826, April 1986.

[105] Hoi Sung Chung, Stefano Piana-Agostinetti, David E Shaw, and William A Eaton. Structural origin of slow diffusion in protein folding. *Science*, 349(6255):1504–1510, 25 September 2015.

[106] M Cioce and A I Lamond. Cajal bodies: a long history of discovery. *Annu. Rev. Cell Dev. Biol.*, 21:105–131, 2005.

[107] Ibrahim I Cisse, Ignacio Izeddin, Sebastien Z Causse, Lydia Boudarene, Adrien Senecal, Leila Muresan, Claire Dugast-Darzacq, Bassam Hajj, Maxime Dahan, and Xavier Darzacq. Real-time dynamics of RNA polymerase II clustering in live human cells. *Science*, 341(6146):664–667, 9 August 2013.

[108] C Clementi and S S Plotkin. The effects of nonnative interactions on protein folding rates: theory and simulation. *Protein Sci.*, 13(7):1750–1766, 2004.

[109] C M Clemson, J N Hutchinson, S A Sara, A W Ensminger, A H Fox, A Chess, and J B Lawrence. An architectural role for a nuclear noncoding RNA: NEAT1 RNA is essential for the structure of paraspeckles. *Mol. Cell*, 33(6):717–726, 2009.

[110] P Cohen. Signal integration at the level of protein kinases, protein phosphatases and their substrates. *Trends Biochem. Sci.*, 17(10):408–413, 1992.

[111] M P Cosma. Daughter-specific repression of saccharomyces cerevisiae HO: Ash1 is the commander. *EMBO Rep.*, 5(10):953–957, 2004.

[112] Susan Costantini, Giovanni Colonna, and Angelo M Facchiano. Amino acid propensities for secondary structures are influenced by the protein structural class. *Biochem. Biophys. Res. Commun.*, 342(2):441–451, 7 April 2006.

[113] J Couthouis, M P Hart, R Erion, O D King, Z Diaz, T Nakaya, F Ibrahim, H J Kim, J Mojsilovic-Petrovic, S Panossian, C E Kim, E C Frackelton, J A Solski, K L Williams, D Clay-Falcone, L Elman, L McCluskey, R Greene, H Hakonarson, R G Kalb, V M Lee, J Q Trojanowski, G A Nicholson, I P Blair, N M Bonini, V M Van Deerlin, Z Mourelatos, J Shorter, and A D Gitler. Evaluating the role of the FUS/TLS-related gene EWSR1 in amyotrophic lateral sclerosis. *Hum. Mol. Genet.*, 21(13):2899–2911, 2012.

[114] Michael David Crabtree, Wade Borcherds, Anusha Poosapati, Sarah L Shammas, Gary W Daughdrill, and Jane Clarke. Conserved helix-flanking prolines modulate IDP: target affinity by altering the lifetime of the bound complex. *Biochemistry*, 2017.

[115] T P Creamer and M N Campbell. Determinants of the polyproline II helix from modeling studies. *Adv. Protein Chem.*, 62:263–282, 2002.

[116] Scott L Crick, Murali Jayaraman, Carl Frieden, Ronald Wetzel, and Rohit V Pappu. Fluorescence correlation spectroscopy shows that monomeric polyglutamine molecules form collapsed structures in aqueous solutions. *Proc. Natl. Acad. Sci. U. S. A.*, 103(45):16764–16769, 7 November 2006.

[117] Scott L Crick, Kiersten M Ruff, Kanchan Garai, Carl Frieden, and Rohit V Pappu. Unmasking the roles of N- and c-terminal flanking sequences from exon 1 of huntingtin as modulators of polyglutamine aggregation. *Proc. Natl. Acad. Sci. U. S. A.*, 110(50):20075–20080, 10 December 2013.

[118] Peter B Crowley and Adel Golovin. Cation-pi interactions in protein-protein interfaces. *Proteins*, 59(2):231–239, 1 May 2005.

[119] Sara Cuylen, Claudia Blaukopf, Antonio Z Politi, Thomas Müller-Reichert, Beate Neumann, Ina Poser, Jan Ellenberg, Anthony A Hyman, and Daniel W Gerlich. Ki-67 acts as a biological surfactant to disperse mitotic chromosomes. *Nature*, 535(7611):308–312, 14 July 2016.

[120] V Daggett, A Li, L S Itzhaki, D E Otzen, and A R Fersht. Structure of the transition state for folding of a protein derived from experiment and simulation. *J. Mol. Biol.*, 257(2):430–440, 29 March 1996.

[121] Valerie Daggett and Alan Fersht. The present view of the mechanism of protein folding. *Nat. Rev. Mol. Cell Biol.*, 4(6):497–502, June 2003.

[122] M E Dahmus. Reversible phosphorylation of the c-terminal domain of RNA polymerase II. *J. Biol. Chem.*, 271(32):19009–19012, 1996.

[123] T Darden, D York, and L Pedersen. Particle mesh ewald: An N· log (n) method for ewald sums in large systems. *J. Chem. Phys.*, 98(12):10089–10092, 1993.

[124] Payel Das, Zhen Xia, and Ruhong Zhou. Collapse of a hydrophobic polymer in a mixture of denaturants. *Langmuir*, 29(15):4877–4882, 2013.

[125] Rahul K Das, Yongqi Huang, Aaron H Phillips, Richard W Kriwacki, and Rohit V Pappu. Cryptic sequence features within the disordered protein p27kip1 regulate cell cycle signaling. *Proc. Natl. Acad. Sci. U. S. A.*, 113(20):5616–5621, 17 May 2016.

[126] Rahul K Das and Rohit V Pappu. Conformations of intrinsically disordered proteins are influenced by linear sequence distributions of oppositely charged residues. *Proceedings of the National Academy of Sciences*, 110(33):13392–13397, 13 August 2013.

[127] Rahul K Das, Kiersten M Ruff, and Rohit V Pappu. Relating sequence encoded information to form and function of intrinsically disordered proteins. *Curr. Opin. Struct. Biol.*, 32(0):102–112, June 2015.

[128] Amrita Dasgupta and Jayant B Udgaonkar. Evidence for initial non-specific polypeptide chain collapse during the refolding of the SH3 domain of PI3 kinase. *J. Mol. Biol.*, 403(3):430–445, 29 October 2010.

[129] Amrita Dasgupta, Jayant B Udgaonkar, and Payel Das. Multistage unfolding of an SH3 domain: An initial Urea-Filled dry molten globule precedes a wet molten globule with Non-Native structure. *J. Phys. Chem. B*, 118(24):6380–6392, 2014.

[130] Norman E Davey and David O Morgan. Building a regulatory network with short linear sequence motifs: Lessons from the degrons of the Anaphase-Promoting complex. *Mol. Cell*, 64(1):12–23, 6 October 2016.

[131] Mansel Davies and Others. *Infra-red spectroscopy and molecular structure*. Elevier Publishing Company, 1963.

[132] Catherine L Day, Callum Smits, F Cindy Fan, Erinna F Lee, W Douglas Fairlie, and Mark G Hinds. Structure of the BH3 domains from the p53-inducible BH3-only proteins noxa and puma in complex with mcl-1. *J. Mol. Biol.*, 380(5):958–971, 25 July 2008.

[133] P.G. de Gennes. *Scaling Concepts in Polymer Physics*. Cornell University Press, 1979.

[134] Carolyn J Decker and Roy Parker. P-bodies and stress granules: possible roles in the control of translation and mRNA degradation. *Cold Spring Harb. Perspect. Biol.*, 4(9):a012286, September 2012.

[135] F Delaglio, S Grzesiek, G W Vuister, G Zhu, J Pfeifer, and A Bax. Nmrpipe - a multidimensional spectral processing system based on unix pipes. *J. Biomol. NMR*, 6(3):277–293, 1995.

[136] A A Deniz, T A Laurence, G S Beligere, M Dahan, A B Martin, D S Chemla, P E Dawson, P G Schultz, and S Weiss. Single-molecule protein folding: diffusion fluorescence resonance energy transfer studies of the denaturation of chymotrypsin inhibitor 2. *Proc. Natl. Acad. Sci. U. S. A.*, 97(10):5179–5184, 2000.

[137] R J Deshaies and J E Ferrell, Jr. Multisite phosphorylation and the countdown to S phase. *Cell*, 107(7):819–822, 2001.

[138] Roger C Diehl, Emily J Guinn, Michael W Capp, Oleg V Tsodikov, and M Thomas Record, Jr. Quantifying additive interactions of the osmolyte proline with individual functional groups of proteins: Comparisons with urea and glycine betaine, interpretation of m-values. *Biochemistry*, 52(35):5997–6010, 3 September 2013.

[139] Manuel Diez, Boris Zimmermann, Michael Börsch, Marcelle König, Enno Schweinberger, Stefan Steigmiller, Rolf Reuter, Suren Felekyan, Volodymyr Kudryavtsev, Claus A M Seidel, and Peter Gräber. Proton-powered subunit rotation in single membrane-bound F0F1-ATP synthase. *Nat. Struct. Mol. Biol.*, 11(2):135–141, 18 January 2004.

[140] Ken Dill and Sarina Bromberg. *Molecular driving forces: statistical thermodynamics in biology, chemistry, physics, and nanoscience*. Garland Science, 2010.

[141] Ken A Dill. Theory for the folding and stability of globular proteins. *Biochemistry*, 24(6):1501–1509, 12 March 1985.

[142] Ken A Dill, Sarina Bromberg, Kaizhi Yue, and Klaus M Fiebig. Principles of protein folding - a perspective from simple exact models. *Protein Sci.*, 4:561–602, 1995.

[143] Ken A Dill and Justin L MacCallum. The protein-folding problem, 50 years on. *Science*, 338(6110):1042–1046, 23 November 2012.

[144] Ruxandra I Dima and D Thirumalai. Asymmetry in the shapes of folded and denatured states of proteins. *J. Phys. Chem. B*, 108(21):6564–6570, 2004.

[145] Mykola Dimura, Thomas O Peulen, Christian A Hanke, Aiswaria Prakash, Holger Gohlke, and Claus Am Seidel. Quantitative FRET studies and integrative modeling unravel the structure and dynamics of biomolecular systems. *Curr. Opin. Struct. Biol.*, 40:163–185, October 2016.

[146] Andrey V Dobrynin and Michael Rubinstein. Flory theory of a polyampholyte chain. *J. Phys. II France*, 5(5):677–695, May 1995.

[147] Christopher M Dobson. Protein folding and misfolding. *Nature*, 426(6968):884–890, 18 December 2003.

[148] Zsuzsanna Dosztányi, Veronika Csizmok, Peter Tompa, and István Simon. IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics*, 21(16):3433–3434, 15 August 2005.

[149] Zsuzsanna Dosztányi, Veronika Csizmók, Péter Tompa, and István Simon. The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. *J. Mol. Biol.*, 347(4):827–839, 8 April 2005.

[150] D A Dougherty. Cation-pi interactions in chemistry and biology: a new view of benzene, phe, tyr, and trp. *Science*, 271(5246):163–168, 12 January 1996.

[151] D A Doyle, J Morais Cabral, R A Pfuetzner, A Kuo, J M Gulbis, S L Cohen, B T Chait, and R MacKinnon. The structure of the potassium channel: molecular basis of k+ conduction and selectivity. *Science*, 280(5360):69–77, 3 April 1998.

[152] Christopher J A Duncan, Siti M B Mohamad, Dan F Young, Andrew J Skelton, T Ronan Leahy, Diane C Munday, Karina M Butler, Sofia Morfopoulou, Julianne R Brown, Mike Hubank, Jeff Connell, Patrick J Gavin, Cathy McMahon, Eugene Dempsey, Niamh E Lynch, Thomas S Jacques, Manoj Valappil, Andrew J Cant, Judith Breuer, Karin R Engelhardt, Richard E Randall, and Sophie Hambleton. Human IFNAR2 deficiency: Lessons for antiviral immunity. *Sci. Transl. Med.*, 7(307):307ra154, 30 September 2015.

[153] A K Dunker, J D Lawson, C J Brown, R M Williams, P Romero, J S Oh, C J Oldfield, A M Campen, C R Ratliff, K W Hipps, J Ausio, M S Nissen, R Reeves, C H Kang, C R Kissinger, R W Bailey, M D Griswold, M Chiu, E C Garner, and Z Obradovic. Intrinsically disordered protein. *J. Mol. Graph. Model.*, 19(1):26–59, 31 May 2001.

[154] P Dunnill. How proteins acquire their structure. *Sci. Prog.*, 53(212):609–619, October 1965.

[155] M Duocastella and C B Arnold. Transient response in ultra-high speed liquid lenses. *Journal of Physics D-Applied Physics*, 46(7), 2013.

[156] Marti Duocastella, Bo Sun, and Craig B Arnold. Simultaneous imaging of multiple focal planes for three-dimensional microscopy using ultra-high-speed adaptive optics. *J. Biomed. Opt.*, 17(5), 2012.

[157] Marti Duocastella, Giuseppe Vicidomini, and Alberto Diaspro. Simultaneous multiplane confocal microscopy using acoustic tunable lenses. *Opt. Express*, 22(16):19293–19301, 11 August 2014.

[158] H Jane Dyson and Peter E Wright. Intrinsically unstructured proteins and their functions. *Nat. Rev. Mol. Cell Biol.*, 6(3):197–208, March 2005.

[159] S J Edelstein. Patterns in the quinary structures of proteins. plasticity and inequivalence of individual molecules in helical arrays of sickle cell hemoglobin and tubulin. *Biophys. J.*, 32(1):347–360, October 1980.

[160] David Eisenberg and Mathias Jucker. The amyloid state of proteins in human diseases. *Cell*, 148(6):1188–1203, 16 March 2012.

[161] W Austin Elam, Travis P Schrank, Andrew J Campagnolo, and Vincent J Hilser. Evolutionary conservation of the polyproline II conformation surrounding intrinsically disordered phosphorylation sites. *Protein Sci.*, 22(4):405–417, April 2013.

[162] Shana Elbaum-Garfinkle, Younghoon Kim, Krzysztof Szczepaniak, Carlos Chih-Hsiung Chen, Christian R Eckmann, Sua Myong, and Clifford P Brangwynne. The disordered P granule protein LAF-1 drives phase separation into droplets with tunable viscosity and dynamics. *Proc. Natl. Acad. Sci. U. S. A.*, 112(23):7189–7194, 9 June 2015.

[163] Elliot L Elson. Fluorescence correlation spectroscopy: past, present, future. *Biophys. J.*, 101(12):2855–2870, 21 December 2011.

[164] Jeremy L England and Gilad Haran. Role of solvation effects in protein denaturation: from thermodynamics to single molecules and back. *Annu. Rev. Phys. Chem.*, 62:257–277, 2011.

[165] Jeremy L England, Vijay S Pande, and Gilad Haran. Chemical denaturants inhibit the onset of dewetting. *J. Am. Chem. Soc.*, 130(36):11854–11855, 2008.

[166] S W Englander. Protein folding intermediates and pathways studied by hydrogen exchange. *Annu. Rev. Biophys. Biomol. Struct.*, 29:213–238, 2000.

[167] E Eyal, R Najmanovich, B J McConkey, M Edelman, and V Sobolev. Importance of solvent accessibility and contact surfaces in modeling side-chain conformations in proteins. *J. Comput. Chem.*, 25(5):712–724, 2004.

[168] P F Faisca, A Nunes, R D Travasso, and E I Shakhnovich. Non-native interactions play an effective role in protein folding dynamics. *Protein Sci.*, 19(11):2196–2209, 2010.

[169] Hanieh Falahati and Eric Wieschaus. Independent active and thermodynamic processes govern the nucleolus assembly in vivo. *Proc. Natl. Acad. Sci. U. S. A.*, 114(6):1335–1340, 7 February 2017.

[170] M Fändrich, M A Fletcher, and C M Dobson. Amyloid fibrils from muscle myoglobin. *Nature*, 410(6825):165–166, 8 March 2001.

[171] M J Fazio, D R Olsen, H Kuivaniemi, M L Chu, J M Davidson, J Rosenbloom, and J Uitto. Isolation and characterization of human elastin cDNAs, and age-associated variation in elastin gene expression in cultured skin fibroblasts. *Lab. Invest.*, 58(3):270–277, 1988.

[172] Marina Feric, Nilesh Vaidya, Tyler S Harmon, Diana M Mitrea, Lian Zhu, Tiffany M Richardson, Richard W Kriwacki, Rohit V Pappu, and Clifford P Brangwynne. Coexisting liquid phases underlie nucleolar subcompartments. *Cell*, 165(7):1686–1697, 16 June 2016.

[173] Diego U Ferreiro, Elizabeth A Komives, and Peter G Wolynes. Frustration in biomolecules. *Q. Rev. Biophys.*, 47(4):285–363, November 2014.

[174] Alan R Fersht. Nucleation mechanisms in protein folding. *Curr. Opin. Struct. Biol.*, 7(1):3–9, 1 February 1997.

[175] Kristen A Fichthorn and W H Weinberg. Theoretical foundations of dynamical monte carlo simulations. *J. Chem. Phys.*, 95(2):1090–1096, 15 July 1991.

[176] Robert D Finn, Alex Bateman, Jody Clements, Penelope Coggill, Ruth Y Eberhardt, Sean R Eddy, Andreas Heger, Kirstie Hetherington, Liisa Holm, Jaina Mistry, Erik L L Sonnhammer, John Tate, and Marco Punta. Pfam: the protein families database. *Nucleic Acids Res.*, 42(Database issue):D222–30, January 2014.

[177] P J Flory. Foundations of rotational isomeric state theory and general methods for generating configurational averages. *Macromolecules*, 7(3):381–392, 1 May 1974.

[178] P J Flory. Introductory lecture. *Faraday Discuss. Chem. Soc.*, 57(0):7–18, 1 January 1974.

[179] Paul J Flory. Thermodynamics of high polymer solutions. *J. Chem. Phys.*, 10(1):51–61, 1 January 1942.

[180] Paul J Flory. *Principles of Polymer Chemistry*. Cornell University Press, Ithaca, NY, 1953.

[181] Paul J Flory. *Statistical Mechanics of Chain Molecules*. Oxford University Press, New York, 1969.

[182] Julie D Forman-Kay and Tanja Mittag. From sequence and forces to structure, function, and evolution of intrinsically disordered proteins. *Structure*, 21(9):1492–1499, 3 September 2013.

[183] Andre Guinier Fornet and Gerard. *Small-Angle Scattering of X-Rays*. Structure of Matter. John Wiley and Sons, New York, 1955.

[184] Archa H Fox, Yun Wah Lam, Anthony K L Leung, Carol E Lyon, Jens Andersen, Matthias Mann, and Angus I Lamond. Paraspeckles: a novel nuclear domain. *Curr. Biol.*, 12(1):13–25, 8 January 2002.

[185] Peter L Freddolino, Christopher B Harrison, Yanxin Liu, and Klaus Schulten. Challenges in protein folding simulations: Timescale, representation, and analysis. *Nat. Phys.*, 6(10):751–758, 1 October 2010.

[186] Peter L Freddolino, Feng Liu, Martin Gruebele, and Klaus Schulten. Ten-microsecond molecular dynamics simulation of a fast-folding WW domain. *Biophys. J.*, 94(10):L75–7, 15 May 2008.

[187] Darón I Freedberg and Philipp Selenko. Live cell NMR. *Annu. Rev. Biophys.*, 43:171–192, 2014.

[188] S A Fromm, J Kamenz, E R Noldeke, A Neu, G Zocher, and R Sprangers. In vitro reconstitution of a cellular phase-transition process that involves the mRNA decapping machinery. *Angew. Chem. Int. Ed Engl.*, 53(28):7354–7359, 2014.

[189] Gustavo Fuertes, Niccolò Banterle, Kiersten M Ruff, Aritra Chowdhury, Davide Mercadante, Christine Koehler, Michael Kachala, Gemma Estrada Girona, Sigrid Milles, Ankur Mishra, Patrick R Onck, Frauke Gräter, Santiago Esteban-Martín, Rohit V Pappu, Dmitri I Svergun, and Edward A Lemke. Decoupling of size and shape fluctuations in heteropolymeric sequences reconciles discrepancies in SAXS vs. FRET measurements. *Proc. Natl. Acad. Sci. U. S. A.*, 114(31):E6342–E6351, 1 August 2017.

[190] Justin P Gallivan and Dennis A Dougherty. Cation-π interactions in structural biology. *Proceedings of the National Academy of Sciences*, 96(17):9459–9464, 17 August 1999.

[191] Justin P Gallivan and Dennis A Dougherty. A computational study of cation- interactions vs salt bridges in aqueous media: Implications for protein engineering. *J. Am. Chem. Soc.*, 122:870–874, 2000.

[192] Basavanapura N Gangadhara, Jennifer M Laine, Sagar V Kathuria, Francesca Massi, and C Robert Matthews. Clusters of branched aliphatic side chains serve as cores of stability in the native state of the HisF TIM barrel protein. *J. Mol. Biol.*, 425(6):1065–1081, 25 March 2013.

[193] Zachary P Gates, Michael C Baxa, Wookyung Yu, Joshua A Riback, Hui Li, Benoît Roux, Stephen B H Kent, and Tobin R Sosnick. Perplexing cooperative folding and stability of a low-sequence complexity, polyproline 2 protein lacking a hydrophobic core. *Proc. Natl. Acad. Sci. U. S. A.*, 114(9):2241–2246, 28 February 2017.

[194] Ehud Gazit. The "correctly folded" state of proteins: is it a metastable state? *Angew. Chem. Int. Ed Engl.*, 41(2):257–259, 18 January 2002.

[195] M B Gee and P E Smith. Kirkwood-Buff theory of molecular and protein association, aggregation, and cellular crowding. *J. Chem. Phys.*, 131(16), 2009.

[196] Lev D Gelb. Monte carlo simulations using sampling from an approximate potential. *J. Chem. Phys.*, 118(17):7747–7750, 1 May 2003.

[197] Y Gerchman, Y Olami, A Rimon, D Taglicht, S Schuldiner, and E Padan. Histidine-226 is part of the ph sensor of NhaA, a Na+/H+ antiporter in escherichia coli. *Proc. Natl. Acad. Sci. U. S. A.*, 90(4):1212–1216, 15 February 1993.

[198] M J Gething and J Sambrook. Protein folding in the cell. *Nature*, 355(6355):33–45, 2 January 1992.

[199] Eric B Gibbs, Feiyue Lu, Bede Portz, Michael J Fisher, Brenda P Medellin, Tatiana N Laremore, Yan Jessie Zhang, David S Gilmour, and Scott A Showalter. Phosphorylation induces sequence-specific conformational switches in the RNA polymerase II c-terminal domain. *Nat. Commun.*, 8:15233, 12 May 2017.

[200] Eric B Gibbs and Scott A Showalter. Quantification of compactness and local order in the ensemble of the intrinsically disordered protein FCP1. *J. Phys. Chem. B*, 120(34):8960–8969, 1 September 2016.

[201] B C Gin, J P Garrahan, and P L Geissler. The limited role of nonnative contacts in the folding pathways of a lattice protein. *J. Mol. Biol.*, 392(5):1303–1314, 2009.

[202] O Glatter and O Kratky. *Small angle x-ray scattering*. Academic Press, London ; New York, 1982.

[203] M Goedert, R Jakes, M G Spillantini, M Hasegawa, M J Smith, and R A Crowther. Assembly of microtubule-associated protein tau into alzheimer-like filaments induced by sulphated glycosaminoglycans. *Nature*, 383(6600):550–553, 10 October 1996.

[204] Rama Reddy Goluguri and Jayant B Udgaonkar. Microsecond rearrangements of hydrophobic clusters in an initially collapsed globule prime structure formation during the folding of a small protein. *J. Mol. Biol.*, 428(15):3102–3117, 2016.

[205] Sven Griep and Uwe Hobohm. PDBselect 1992-2009 and PDBfilter-select. *Nucleic Acids Res.*, 38(Database issue):D318–D319, January 2010.

[206] A Yu Grosberg and D V Kuznetsov. Quantitative theory of the globule-to-coil transition. 1. link density distribution in a globule and its radius of gyration. *Macromolecules*, 25(7):1970–1979, 1 March 1992.

[207] Alan Grossfield and Daniel M Zuckerman. Chapter 2 quantifying uncertainty and sampling quality in biomolecular simulations. In Ralph A Wheeler, editor, *Annual Reports in Computational Chemistry*, volume Volume 5, pages 23–48. Elsevier, 2009.

[208] Dominika T Gruszka, Justyna A Wojdyla, Richard J Bingham, Johan P Turkenburg, Iain W Manfield, Annette Steward, Andrew P Leech, Joan A Geoghegan, Timothy J Foster, Jane Clarke, and Others. Staphylococcal biofilm-forming protein has a contiguous rod-like structure. *Proceedings of the National Academy of Sciences*, 109(17):E1011–E1018, 2012.

[209] P J Gualfetti, M Iwakura, J C Lee, H Kihara, O Bilsel, J A Zitzewitz, and C R Matthews. Apparent radii of the native, stable intermediates and unfolded conformers of the alpha-subunit of tryptophan synthase from e. coli, a TIM barrel protein. *Biochemistry*, 38(40):13367–13378, 1999.

[210] Nicolas Guex and Manuel C Peitsch. SWISS-MODEL and the Swiss-Pdb viewer: an environment for comparative protein modeling. *Electrophoresis*, 18(15):2714–2723, 1997.

[211] E J Guinn, J J Schwinefus, H K Cha, J L McDevitt, W E Merker, R Ritzer, G W Muth, S W Engelsgjerd, K E Mangold, P J Thompson, M J Kerins, and M T Record. Quantifying functional group interactions that determine urea effects on nucleic acid helix formation (vol 135, pg 5828, 2013). *J. Am. Chem. Soc.*, 135(24):9220–9220, 2013.

[212] Emily J Guinn, Laurel M Pegram, Michael W Capp, Michelle N Pollock, and M Thomas Record. Quantifying why urea is a protein denaturant, whereas glycine betaine is a protein stabilizer. *Proceedings of the National Academy of Sciences*, 108(41):16932–16937, 19 September 2011.

[213] Jeremy Gunawardena. Multisite protein phosphorylation makes a good threshold but can be a poor switch. *Proc. Natl. Acad. Sci. U. S. A.*, 102(41):14617–14622, 11 October 2005.

[214] Z Guo and D Thirumalai. The nucleation-collapse mechanism in protein folding: evidence for the non-uniqueness of the folding nucleus. *Fold. Des.*, 2(6):377–391, 1997.

[215] Alex Gutteridge and Janet M Thornton. Understanding nature's catalytic toolkit. *Trends Biochem. Sci.*, 30(11):622–629, November 2005.

[216] B Y Ha and D Thirumalai. Semiflexible chains under tension. *J. Chem. Phys.*, 106(10):4243–4247, 1997.

[217] Tina W Han, Masato Kato, Shanhai Xie, Leeju C Wu, Hamid Mirzaei, Jimin Pei, Min Chen, Yang Xie, Jeffrey Allen, Guanghua Xiao, and Steven L McKnight. Cell-free

formation of RNA granules: bound RNAs identify features and components of cellular assemblies. *Cell*, 149(4):768–779, 5 November 2012.

[218] Momoyo Hanazawa, Masafumi Yonetani, and Asako Sugimoto. PGL proteins self associate and bind RNPs to mediate germ granule assembly in c. elegans. *J. Cell Biol.*, 192(6):929–937, 21 March 2011.

[219] Gilad Haran. How, when and why proteins collapse: the relation to folding. *Curr. Opin. Struct. Biol.*, 22(1):14–20, 19 November 2011.

[220] S E Harding and P Johnson. The concentration-dependence of macromolecular parameters. *Biochem. J*, 231(3):543–547, 1 November 1985.

[221] Tyler S Harmon, Michael D Crabtree, Sarah L Shammas, Ammon E Posey, Jane Clarke, and Rohit V Pappu. GADIS: Algorithm for designing sequences to achieve target secondary structure profiles of intrinsically disordered proteins. *Protein Eng. Des. Sel.*, 29(9):339–346, September 2016.

[222] Tyler S Harmon, Alex S Holehouse, and Rohit V Pappu. To mix, or to demix, that is the question. *Biophys. J.*, 112(4):565–567, 28 February 2017.

[223] Tyler S Harmon, Alex S Holehouse, Michael K Rosen, and Rohit V Pappu. Intrinsically disordered linkers determine the interplay between phase separation and gelation in multivalent proteins. 16 July 2017.

[224] Y Harpaz, M Gerstein, and C Chothia. Volume changes on protein folding. *Structure*, 2(7):641–649, 15 July 1994.

[225] Thomas K Harris and George J Turner. Structural basis of perturbed pka values of catalytic groups in enzyme active sites. *IUBMB Life*, 53(2):85–98, February 2002.

[226] Wafa Hassouneh, Ekaterina B Zhulina, Ashutosh Chilkoti, and Michael Rubinstein. Elastin-like polypeptide diblock copolymers Self-Assemble into weak micelles. *Macromolecules*, 48(12):4183–4195, 2015.

[227] W K Hastings. Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.

[228] Elke Haustein and Petra Schwille. Fluorescence correlation spectroscopy: Novel variations of an established technique. *Annu. Rev. Biophys. Biomol. Struct.*, 36(1):151–169, 2007.

[229] Katherine A Henzler-Wildman, Ming Lei, Vu Thai, S Jordan Kerns, Martin Karplus, and Dorothee Kern. A hierarchy of timescales in protein dynamics is linked to enzyme catalysis. *Nature*, 450(7171):913–916, 6 December 2007.

[230] Berk Hess, Henk Bekker, Herman J C Berendsen, and Johannes G E M Fraaije. LINCS: A linear constraint solver for molecular simulations. *J. Comput. Chem.*, 18(12):1463–1472, 1997.

[231] A Hiroki, Y Maekawa, M Yoshida, K Kubota, and R Katakai. Volume phase transitions of poly(acryloyl-l-proline methyl ester) gels in response to water–alcohol composition. *Polymer*, 42(5):1863–1867, 2001.

[232] Denes Hnisz, Krishna Shrinivas, Richard A Young, Arup K Chakraborty, and Phillip A Sharp. A phase separation model for transcriptional control. *Cell*, 169(1):13–23, 23 March 2017.

[233] D W Hoffman, C Davies, S E Gerchman, J H Kycia, S J Porter, S W White, and V Ramakrishnan. Crystal structure of prokaryotic ribosomal protein l9: a bi-lobed RNA-binding protein. *EMBO J.*, 13(1):205–212, 1994.

[234] Hagen Hofmann, Andrea Soranno, Alessandro Borgia, Klaus Gast, Daniel Nettels, and Benjamin Schuler. Polymer scaling laws of unfolded and intrinsically disordered proteins quantified with single-molecule spectroscopy. *Proc. Natl. Acad. Sci. U. S. A.*, 109(40):16155–16160, 14 September 2012.

[235] Alex S Holehouse, Rahul K Das, James N Ahad, Mary O G Richardson, and Rohit V Pappu. CIDER: Resources to analyze Sequence-Ensemble relationships of intrinsically disordered proteins. *Biophys. J.*, 112(1):16–21, 10 January 2017.

[236] Alex S Holehouse, Kanchan Garai, Nicholas Lyle, Andreas Vitalis, and Rohit V Pappu. Quantitative assessments of the distinct contributions of polypeptide backbone amides versus side chain groups to chain expansion via chemical denaturation. *J. Am. Chem. Soc.*, 137(8):2984–2995, 4 March 2015.

[237] Alex S Holehouse and Kristen M Naegle. Reproducible analysis of Post-Translational modifications in Proteomes–Application to human mutations. *PLoS One*, 10(12):e0144692, 14 December 2015.

[238] C I Holmberg, S E Tran, J E Eriksson, and L Sistonen. Multisite phosphorylation provides sophisticated regulation of transcription factors. *Trends Biochem. Sci.*, 27(12):619–627, 2002.

[239] L M F Holthauzen, J Rosgen, and D W Bolen. Hydrogen bonding progressively strengthens upon transfer of the protein Urea-Denatured state to water and protecting osmolytes. *Biochemistry*, 49(6):1310–1318, 2010.

[240] Wolf Holtkamp, Goran Kokic, Marcus Jäger, Joerg Mittelstaet, Anton A Komar, and Marina V Rodnina. Cotranslational protein folding on the ribosome monitored in real time. *Science*, 350(6264):1104–1107, 27 November 2015.

[241] Dominik Horinek and Roland R Netz. Can simulations quantitatively predict peptide transfer free energies to urea solutions? thermodynamic concepts and force field limitations. *J. Phys. Chem. A*, 115(23):6125–6136, 1 March 2011.

[242] Yang Hsia, Jacob B Bale, Shane Gonen, Dan Shi, William Sheffler, Kimberly K Fong, Una Nattermann, Chunfu Xu, Po-Ssu Huang, Rashmi Ravichandran, Sue Yi, Trisha N Davis, Tamir Gonen, Neil P King, and David Baker. Design of a hyperstable 60-subunit protein dodecahedron. [corrected]. *Nature*, 535(7610):136–139, 7 July 2016.

[243] Lan Hua, Ruhong Zhou, D Thirumalai, and B J Berne. Urea denaturation by stronger dispersion interactions with proteins than water implies a 2-stage unfolding. *Proceedings of the National Academy of Sciences*, 105(44):16928–16933, 4 November 2008.

[244] F Huang, L Ying, and A R Fersht. Direct observation of barrier-limited folding of BBL by single-molecule fluorescence resonance energy transfer. *Proc. Natl. Acad. Sci. U. S. A.*, 106(38):16239–16244, 2009.

[245] Jing Huang, Sarah Rauscher, Grzegorz Nawrocki, Ting Ran, Michael Feig, Bert L de Groot, Helmut Grubmüller, and Alexander D MacKerell, Jr. CHARMM36m: an improved force field for folded and intrinsically disordered proteins. *Nat. Methods*, 14(1):71–73, January 2017.

[246] M L Huggins. Solutions of long chain compounds. *J. Chem. Phys.*, 9(5):440–440, 1941.

[247] Maurice L Huggins. Some properties of solutions of long-chain compounds. *J. Phys. Chem.*, 46(1):151–158, 1 January 1942.

[248] Michael P Hughes, Michael R Sawaya, Lukasz Goldschmidt, Jose A Rodriguez, Duilio Cascio, Tamir Gonen, and David S Eisenberg. Low-complexity domains adhere by reversible amyloid-like interactions between kinked β-sheets. 22 June 2017.

[249] Greta Hultqvist, Emma Åberg, Carlo Camilloni, Gustav N Sundell, Eva Andersson, Jakob Dogan, Celestine N Chi, Michele Vendruscolo, and Perk Jemth. Emergence and evolution of an interaction between intrinsically disordered proteins. *Elife*, 6, 11 April 2017.

[250] W Humphrey, A Dalke, and K Schulten. VMD: Visual molecular dynamics. *J. Mol. Graph. Model.*, 14(1):33–8, 27–8, February 1996.

[251] Anthony A Hyman, Christoph A Weber, and Frank Jülicher. Liquid-liquid phase separation in biology. *Annu. Rev. Cell Dev. Biol.*, 30:39–58, 2014.

[252] Lilia M Iakoucheva, Predrag Radivojac, Celeste J Brown, Timothy R O'Connor, Jason G Sikes, Zoran Obradovic, and A Keith Dunker. The importance of intrinsic disorder for protein phosphorylation. *Nucleic Acids Res.*, 32(3):1037–1049, 11 February 2004.

[253] Barbara Jachimska, Monika Wasilewska, and Zbigniew Adamczyk. Characterization of globular protein solutions by dynamic light scattering, electrophoretic mobility, and viscosity measurements. *Langmuir*, 24(13):6866–6872, 1 June 2008.

[254] S E Jackson. How do small single-domain proteins fold? *Fold. Des.*, 3(4):R81–91, 1998.

[255] Jaby Jacob, Bryan Krantz, Robin S Dothager, P Thiyagarajan, and Tobin R Sosnick. Early collapse is not an obligate step in protein folding. *J. Mol. Biol.*, 338(2):369–382, 23 April 2004.

[256] William M Jacobs and Daan Frenkel. Predicting phase behavior in multicomponent mixtures. *J. Chem. Phys.*, 139(2):024108, 14 July 2013.

[257] William M Jacobs and Daan Frenkel. Phase transitions in biological systems with many components. *Biophys. J.*, 112(4):683–691, 28 February 2017.

[258] G N Jacobson and P L Clark. Quality over quantity: optimizing co-translational protein folding with non-'optimal'synonymous codons. *Curr. Opin. Struct. Biol.*, 2016.

[259] T R Jahn and S E Radford. Folding versus aggregation: polypeptide conformations on competing pathways. *Arch. Biochem. Biophys.*, 469(1):100–117, 2008.

[260] Ankur Jain and Ronald D Vale. RNA phase transitions in repeat expansion disorders. *Nature*, 546(7657):243–247, 8 June 2017.

[261] Saumya Jain, Joshua R Wheeler, Robert W Walters, Anurag Agrawal, Anthony Barsic, and Roy Parker. ATPase-Modulated stress granules contain a diverse proteome and substructure. *Cell*, 164(3):487–498, 28 January 2016.

[262] P A Jennings and P E Wright. Formation of a molten globule intermediate early in the kinetic folding pathway of apomyoglobin. *Science*, 262(5135):892–896, 5 November 1993.

[263] Malene Ringkjøbing Jensen, Rob W H Ruigrok, and Martin Blackledge. Describing intrinsically disordered proteins at atomic resolution by NMR. *Curr. Opin. Struct. Biol.*, 23(3):426–435, June 2013.

[264] Abhishek K Jha, Andrés Colubri, Karl F Freed, and Tobin R Sosnick. Statistical coil model of the unfolded state: Resolving the reconciliation problem. *Proc. Natl. Acad. Sci. U. S. A.*, 102(37):13099–13104, 13 September 2005.

[265] Santosh Kumar Jha and Susan Marqusee. Kinetic evidence for a two-stage mechanism of protein denaturation by guanidinium chloride. *Proc. Natl. Acad. Sci. U. S. A.*, 111(13):4856–4861, 17 March 2014.

[266] Hao Jiang, Shusheng Wang, Yuejia Huang, Xiaonan He, Honggang Cui, Xueliang Zhu, and Yixian Zheng. Phase transition of spindle-associated protein regulate spindle apparatus assembly. *Cell*, 163(1):108–122, 24 September 2015.

[267] Kai Jiang, Ce Zhang, Durgarao Guttula, Fan Liu, Jeroen A van Kan, Christophe Lavelle, Krzysztof Kubiak, Antoine Malabirade, Alain Lapp, Véronique Arluison, and Johan R C van der Maarel. Effects of hfq on the conformation and compaction of DNA. *Nucleic Acids Res.*, 43(8):4332–4341, 30 April 2015.

[268] José L Jiménez, Ewan J Nettleton, Mario Bouchard, Carol V Robinson, Christopher M Dobson, and Helen R Saibil. The protofilament structure of insulin amyloid fibrils. *Proc. Natl. Acad. Sci. U. S. A.*, 99(14):9196–9201, 9 July 2002.

[269] N Jones, I M Blasutig, V Eremina, J M Ruston, F Bladt, H Li, H Huang, L Larose, S S Li, T Takano, S E Quaggin, and T Pawson. Nck adaptor proteins link nephrin to the actin cytoskeleton of kidney podocytes. *Nature*, 440(7085):818–823, 2006.

[270] N Jones, L A New, M A Fortino, V Eremina, J Ruston, I M Blasutig, L Aoudjit, Y Zou, X Liu, G L Yu, T Takano, S E Quaggin, and T Pawson. Nck proteins maintain the adult glomerular filtration barrier. *J. Am. Soc. Nephrol.*, 20(7):1533–1543, 2009.

[271] S Jones and J M Thornton. Principles of protein-protein interactions. *Proc. Natl. Acad. Sci. U. S. A.*, 93(1):13–20, 9 January 1996.

[272] Paul Jorgensen, Joy L Nishikawa, Bobby-Joe Breitkreutz, and Mike Tyers. Systematic identification of pathways that couple cell growth and division in yeast. *Science*, 297(5580):395–400, 19 July 2002.

[273] William L Jorgensen, Jayaraman Chandrasekhar, Jeffry D Madura, Roger W Impey, and Michael L Klein. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.*, 79(2):926–935, 15 July 1983.

[274] Wolfgang Kabsch and Christian Sander. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22(12):2577–2637, 1983.

[275] Laxmikant Kalé, Robert Skeel, Milind Bhandarkar, Robert Brunner, Attila Gursoy, Neal Krawetz, James Phillips, Aritomo Shinozaki, Krishnan Varadarajan, and Klaus Schulten. NAMD2: Greater scalability for parallel molecular dynamics. *J. Comput. Phys.*, 151(1):283–312, 1 May 1999.

[276] George A Kaminski, Richard A Friesner, Julian Tirado-Rives, and William L Jorgensen. Evaluation and reparametrization of the OPLS-AA force field for proteins via comparison with accurate quantum chemical calculations on peptides. *J. Phys. Chem. B*, 105(28):6474–6487, 2001.

[277] Hongsuk Kang, Francisco X Vázquez, Leili Zhang, Payel Das, Leticia Marisel Toledo-Sherman, Binquan Luan, Michael Levitt, and Ruhong Zhou. Emerging β-sheet rich conformations in super-compact huntingtin exon-1 mutant structures. *J. Am. Chem. Soc.*, 2017.

[278] Evgeny Kanshin, Louis-Philippe Bergeron-Sandoval, S Sinan Isik, Pierre Thibault, and Stephen W Michnick. A cell-signaling network temporally resolves specific versus promiscuous phosphorylation. *Cell Rep.*, 10(7):1202–1214, 24 February 2015.

[279] M Karplus and D L Weaver. Protein folding dynamics: the diffusion-collision model and experimental data. *Protein Sci.*, 3(4):650–668, April 1994.

[280] Martin Karplus and David L Weaver. Protein-folding dynamics. *Nature*, 260(5550):404–406, 1 April 1976.

[281] Sagar V Kathuria, Yvonne H Chan, R Paul Nobrega, Ayşegül Özen, and C Robert Matthews. Clusters of isoleucine, leucine, and valine side chains define cores of stability in high-energy states of globular proteins: Sequence determinants of structure and stability. *Protein Sci.*, 25(3):662–675, March 2016.

[282] Masato Kato, Tina W Han, Shanhai Xie, Kevin Shi, Xinlin Du, Leeju C Wu, Hamid Mirzaei, Elizabeth J Goldsmith, Jamie Longgood, Jimin Pei, Nick V Grishin, Douglas E Frantz, Jay W Schneider, She Chen, Lin Li, Michael R Sawaya, David Eisenberg, Robert Tycko, and Steven L McKnight. Cell-free formation of RNA granules: low complexity sequence domains form dynamic fibers within hydrogels. *Cell*, 149(4):753–767, 11 May 2012.

[283] Sarah K Kaufman, David W Sanders, Talitha L Thomas, Allison J Ruchinskas, Jaime Vaquer-Alicea, Apurwa M Sharma, Timothy M Miller, and Marc I Diamond. Tau prion strains dictate patterns of cell pathology, progression rate, and regional vulnerability in vivo. *Neuron*, 92(4):796–812, 23 November 2016.

[284] Rochus L J Keller. *The Computer Aided Resonance Assignment Tutorial*. CANTINA Verlag, Goldau, Swizterland, 1 edition, 2004.

[285] J C Kendrew, G Bodo, H M Dintzis, R G Parrish, H Wyckoff, and D C Phillips. A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. *Nature*, 181(4610):662–666, 8 March 1958.

[286] Ashish K Khandpur, Stephan Foerster, Frank S Bates, Ian W Hamley, Anthony J Ryan, Wim Bras, Kristoffer Almdal, and Kell Mortensen. Polyisoprene-Polystyrene diblock copolymer phase diagram near the Order-Disorder transition. *Macromolecules*, 28(26):8796–8806, 1 December 1995.

[287] Alexey G Kikhney and Dmitri I Svergun. A practical guide to small angle x-ray scattering (SAXS) of flexible and intrinsically disordered proteins. *FEBS Lett.*, 29 August 2015.

[288] Peter S Kim and Robert L Baldwin. Specific intermediates in the folding reactions of small proteins and the mechanism of protein folding. *Annu. Rev. Biochem.*, 51(1):459–489, 1 June 1982.

[289] Sangsik Kim, Jun Huang, Yongjin Lee, Sandipan Dutta, Hee Young Yoo, Young Mee Jung, Yongseok Jho, Hongbo Zeng, and Dong Soo Hwang. Complexation and coacervation of like-charged polyelectrolytes inspired by mussels. *Proc. Natl. Acad. Sci. U. S. A.*, 113(7):E847–53, 16 February 2016.

[290] Yujin E Kim, Mark S Hipp, Andreas Bracher, Manajit Hayer-Hartl, and F Ulrich Hartl. Molecular chaperone functions in protein folding and proteostasis. *Annu. Rev. Biochem.*, 82:323–355, 2013.

[291] T Kimura, T Uzawa, K Ishimori, I Morishima, S Takahashi, T Konno, S Akiyama, and T Fujisawa. Specific collapse followed by slow hydrogen-bond formation of beta-sheet in the folding of single-chain monellin. *Proc. Natl. Acad. Sci. U. S. A.*, 102(8):2748–2753, 2005.

[292] Ebru Kizilay, A Basak Kayitmazer, and Paul L Dubin. Complexation and coacervation of polyelectrolytes with oppositely charged colloids. *Adv. Colloid Interface Sci.*, 167(1-2):24–37, 14 September 2011.

[293] J Klein-Seetharaman, M Oikawa, S B Grimshaw, J Wirmer, E Duchardt, T Ueda, T Imoto, L J Smith, C M Dobson, and H Schwalbe. Long-range interactions within a nonnative protein. *Science*, 295(5560):1719–1722, 2002.

[294] John L Klepeis, Kresten Lindorff-Larsen, Ron O Dror, and David E Shaw. Long-timescale molecular dynamics simulations of protein structure and function. *Curr. Opin. Struct. Biol.*, 19(2):120–127, April 2009.

[295] D K Klimov and D Thirumalai. Criterion that determines the foldability of proteins. *Phys. Rev. Lett.*, 76(21):4070–4073, 1996.

[296] Tuomas P J Knowles, Michele Vendruscolo, and Christopher M Dobson. The amyloid state and its association with protein misfolding diseases. *Nat. Rev. Mol. Cell Biol.*, 15(6):384–396, June 2014.

[297] Jonathan E Kohn, Ian S Millett, Jaby Jacob, Bojan Zagrovic, Thomas M Dillon, Nikolina Cingel, Robin S Dothager, Soenke Seifert, P Thiyagarajan, Tobin R Sosnick, M Zahid Hasan, Vijay S Pande, Ingo Ruczinski, Sebastian Doniach, and Kevin W Plaxco. Random-coil behavior and the dimensions of chemically unfolded proteins.

Proceedings of the National Academy of Sciences, 101(34):12491–12496, 16 August 2004.

[298] M Koivomagi, M Ord, A Iofik, E Valk, R Venta, I Faustova, R Kivi, E R M Balog, S M Rubin, and M Loog. Multisite phosphorylation networks as signal processors for cdk1. *Nat. Struct. Mol. Biol.*, 20(12):1415–1424, 2013.

[299] Hironori Kokubo and B Montgomery Pettitt. Preferential solvation in urea solutions at different concentrations: properties from simulation studies. *J. Phys. Chem. B*, 111(19):5233–5242, 21 April 2007.

[300] P V Konarev, V V Volkov, A V Sokolova, M H J Koch, and D I Svergun. PRIMUS: a windows PC-based system for small-angle scattering data analysis. *J. Appl. Crystallogr.*, 36:1277–1282, 2003.

[301] Iwo König, Arash Zarrine-Afsar, Mikayel Aznauryan, Andrea Soranno, Bengt Wunderlich, Fabian Dingfelder, Jakob C Stüber, Andreas Plückthun, Daniel Nettels, and Benjamin Schuler. Single-molecule spectroscopy of protein conformational dynamics in live eukaryotic cells. *Nat. Methods*, 12(8):773–779, August 2015.

[302] Robert Konrat. NMR contributions to structural dynamics studies of intrinsically disordered proteins. *J. Magn. Reson.*, 241:74–85, April 2014.

[303] David A Korasick, Corey S Westfall, Soon Goo Lee, Max H Nanao, Renaud Dumas, Gretchen Hagen, Thomas J Guilfoyle, Joseph M Jez, and Lucia C Strader. Molecular basis for AUXIN RESPONSE FACTOR protein interaction and the control of auxin response repression. *Proc. Natl. Acad. Sci. U. S. A.*, 111(14):5427–5432, 8 April 2014.

[304] Denes Kovacs and Peter Tompa. Diverse functional manifestations of intrinsic structural disorder in molecular chaperones. *Biochem. Soc. Trans.*, 40(5):963–968, October 2012.

[305] H A Kramers. Brownian motion in a field of force and the diffusion model of chemical reactions. *Physica*, 7(4):284–304, 1 April 1940.

[306] Oleg Krichevsky and Grégoire Bonnet. Fluorescence correlation spectroscopy: the technique and its applications. *Rep. Prog. Phys.*, 65(2):251, 2002.

[307] Sonja Kroschwald, Shovamayee Maharana, Daniel Mateju, Liliana Malinovska, Elisabeth Nüske, Ina Poser, Doris Richter, and Simon Alberti. Promiscuous interactions and protein disaggregases determine the material state of stress-inducible RNP granules. *Elife*, 4:e06807, 4 August 2015.

[308] B Kuhlman, D L Luisi, P A Evans, and D P Raleigh. Global analysis of the effects of temperature and denaturant on the folding and unfolding kinetics of the n-terminal domain of the protein L9. *J. Mol. Biol.*, 284(5):1661–1670, 1998.

[309] Praveen Kumar, Michael S Chimenti, Hayley Pemble, Andre Schonichen, Oliver Thompson, Matthew P Jacobson, and Torsten Wittmann. Multisite phosphorylation disrupts arginine-glutamate salt bridge networks required for binding of cytoplasmic linker-associated protein 2 (CLASP2) to end-binding protein 1 (EB1). *J. Biol. Chem.*, 287(21):17050–17064, 18 May 2012.

[310] S Kumar and R Nussinov. Salt bridge stability in monomeric proteins. *J. Mol. Biol.*, 293(5):1241–1255, 12 November 1999.

[311] Sanjay Kumar and Jan H Hoh. Modulation of repulsive forces between neurofilaments by sidearm phosphorylation. *Biochem. Biophys. Res. Commun.*, 324(2):489–496, 12 November 2004.

[312] Ilmin Kwon, Masato Kato, Siheng Xiang, Leeju Wu, Pano Theodoropoulos, Hamid Mirzaei, Tina Han, Shanhai Xie, Jeffry L Corden, and Steven L McKnight. Phosphorylation-regulated binding of RNA polymerase II to fibrous polymers of low-complexity domains. *Cell*, 155(5):1049–1060, 21 November 2013.

[313] Ilmin Kwon, Siheng Xiang, Masato Kato, Leeju Wu, Pano Theodoropoulos, Tao Wang, Jiwoong Kim, Jonghyun Yun, Yang Xie, and Steven L McKnight. Poly-dipeptides encoded by the c9orf72 repeats bind nucleoli, impede RNA biogenesis, and kill cells. *Science*, 345(6201):1139–1145, 5 September 2014.

[314] Valérie Lallemand-Breitenbach and Hugues de Thé. PML nuclear bodies. *Cold Spring Harb. Perspect. Biol.*, 2(5):a000661, May 2010.

[315] Angus I Lamond and David L Spector. Nuclear speckles: a model for nuclear organelles. *Nat. Rev. Mol. Cell Biol.*, 4(8):605–612, August 2003.

[316] K Lang, F X Schmid, and G Fischer. Catalysis of protein folding by prolyl isomerase. *Nature*, 329(6136):268–270, 1987.

[317] L J Lapidus, W A Eaton, and J Hofrichter. Measuring the rate of intramolecular contact formation in polypeptides. *Proc. Natl. Acad. Sci. U. S. A.*, 97(13):7220–7225, 20 June 2000.

[318] Andrew R Leach. *Molecular modelling: principles and applications*. Pearson education, 2001.

[319] Changhwan Lee, Patricia Occhipinti, and Amy S Gladfelter. PolyQ-dependent RNA-protein assemblies control symmetry breaking. *J. Cell Biol.*, 208(5):533–544, 2 March 2015.

[320] Chiu Fan Lee, Clifford P Brangwynne, Joebin Gharakhani, Anthony A Hyman, and Frank Juelicher. Spatial organization of the cell cytoplasm by position-dependent phase separation. *Phys. Rev. Lett.*, 111(8), 2013.

[321] A Lempel and J Ziv. On the complexity of finite sequences. *IEEE Trans. Inf. Theory*, 22(1):75–81, January 1976.

[322] A M Lesk and C Chothia. How different amino acid sequences determine similar protein structures: the structure and evolutionary dynamics of the globins. *J. Mol. Biol.*, 136(3):225–270, 25 January 1980.

[323] J F Leszczynski and G D Rose. Loops in globular proteins: a novel category of secondary structure. *Science*, 234(4778):849–855, 14 November 1986.

[324] Hoi Tik Alvin Leung, Olivier Bignucolo, Regula Aregger, Sonja A Dames, Adam Mazur, Simon Bernèche, and Stephan Grzesiek. A rigorous and efficient method to reweight very large conformational ensembles using average experimental data and to determine their relative information content. *J. Chem. Theory Comput.*, 12(1):383–394, 2016.

[325] Matteo Levantino, Briony A Yorke, Diana Cf Monteiro, Marco Cammarata, and Arwen R Pearson. Using synchrotrons and XFELs for time-resolved x-ray crystallography and solution scattering experiments on biomolecules. *Curr. Opin. Struct. Biol.*, 35:41–48, December 2015.

[326] Cyrus Levinthal. How to fold graciously. In J T P DeBrunner and E Munck, editors, *Mossbauer Spectroscopy in Biological Systems*, pages 22–24. University of Illinois Press, 1969.

[327] M Levitt and S Lifson. Refinement of protein conformations using a macromolecular energy minimization procedure. *J. Mol. Biol.*, 46(2):269–279, 14 December 1969.

[328] Daniel Lhuillier. A simple model for polymeric fractals in a good solvent and an improved version of the flory approximation. *J. Phys. France*, 49(5):705–710, 1 May 1988.

[329] Nan K Li, William H Fuss, Lei Tang, Renpeng Gu, Ashutosh Chilkoti, Stefan Zauscher, and Yaroslava G Yingling. Prediction of solvent-induced morphological changes of polyelectrolyte diblock copolymer micelles. *Soft Matter*, 11(42):8236–8245, 18 August 2015.

[330] Pilong Li, Sudeep Banjade, Hui-Chun Cheng, Soyeon Kim, Baoyu Chen, Liang Guo, Marc Llaguno, Javoris V Hollingsworth, David S King, Salman F Banani, Paul S Russo, Qiu-Xing Jiang, B Tracy Nixon, and Michael K Rosen. Phase transitions in the assembly of multivalent signalling proteins. *Nature*, 483(7389):336–340, 15 March 2012.

[331] Xinyan Li, Song Jiang, and Richard I Tapping. Toll-like receptor signaling in cell proliferation and survival. *Cytokine*, 49(1):1–9, January 2010.

[332] I M Lifshitz and V V Slyozov. The kinetics of precipitation from supersaturated solid solutions. *J. Phys. Chem. Solids*, 19(1):35–50, 1 April 1961.

[333] Woon Ki Lim, Joerg Roesgen, and S Walter Englander. Urea, but not guanidinium, destabilizes proteins by forming hydrogen bonds to the peptide group. *Proc. Natl. Acad. Sci. U. S. A.*, 106(8):2595–2600, 24 February 2009.

[334] Milo M Lin and Ahmed H Zewail. Hydrophobic forces and the length limit of foldable protein domains. *Proc. Natl. Acad. Sci. U. S. A.*, 109(25):9851–9856, 19 June 2012.

[335] Yi-Hsuan Lin and Hue Sun Chan. Phase separation and Single-Chain compactness of charged disordered proteins are strongly correlated. *Biophys. J.*, 0(0), 5 May 2017.

[336] Yi-Hsuan Lin, Julie D Forman-Kay, and Hue Sun Chan. Sequence-Specific polyampholyte phase separation in membraneless organelles. *Phys. Rev. Lett.*, 117(17):178101, 21 October 2016.

[337] Yi-Hsuan Lin, Jianhui Song, Julie D Forman-Kay, and Hue Sun Chan. Random-phase-approximation theory for sequence-dependent, biologically functional liquid-liquid phase separation of intrinsically disordered proteins. *J. Mol. Liq.*, 2016.

[338] Yuan Lin, David S W Protter, Michael K Rosen, and Roy Parker. Formation and maturation of phase-separated liquid droplets by RNA-binding proteins. *Mol. Cell*, 60(2):208–219, 15 October 2015.

[339] Kresten Lindorff-Larsen, Stefano Piana, Ron O Dror, and David E Shaw. How fast-folding proteins fold. *Science*, 334(6055):517–520, 28 October 2011.

[340] Shuo-Chien Ling, Magdalini Polymenidou, and Don W Cleveland. Converging mechanisms in ALS and FTD: disrupted RNA and protein homeostasis. *Neuron*, 79(3):416–438, 7 August 2013.

[341] Q Liu, B Larsen, M Ricicova, S Orlicky, H Tekotte, X Tang, K Craig, A Quiring, T Le Bihan, C Hansen, F Sicheri, and M Tyers. SCFCdc4 enables mating type switching in yeast by cyclin-dependent kinase-mediated elimination of the ash1 transcriptional repressor. *Mol. Cell. Biol.*, 31(3):584–598, 2011.

[342] Zhenxing Liu, Govardhan Reddy, Edward P O'Brien, and D Thirumalai. Collapse kinetics and chevron plots from simulations of denaturant-dependent folding of globular proteins. *Proc. Natl. Acad. Sci. U. S. A.*, 108(19):7787–7792, 10 May 2011.

[343] Harvey Lodish, Arnold Berk, S Lawrence Zipursky, Paul Matsudaira, David Baltimore, and James Darnell. *Molecular Cell Biology*. W. H. Freeman, 2000.

[344] T M Lohman, L B Overman, M E Ferrari, and A G Kozlov. A highly salt-dependent enthalpy change for escherichia coli SSB protein-nucleic acid binding due to ion-protein interactions. *Biochemistry*, 35(16):5272–5279, 23 April 1996.

[345] K P Lu, Y C Liou, and X Z Zhou. Pinning down proline-directed phosphorylation signaling. *Trends Cell Biol.*, 12(4):164–172, 2002.

[346] Xiaomeng Lu and Regina M Murphy. Asparagine repeat peptides: Aggregation kinetics and comparison with glutamine repeats. *Biochemistry*, 54(31):4784–4794, 11 August 2015.

[347] Bowu Luan, Nicholas Lyle, Rohit V Pappu, and Daniel P Raleigh. Denatured state ensembles with the same radii of gyration can form significantly different long-range contacts. *Biochemistry*, 53(1):39–47, 14 January 2014.

[348] Nicholas Lyle, Rahul K Das, and Rohit V Pappu. A quantitative measure for protein conformational heterogeneity. *J. Chem. Phys.*, 139(12):09B607_1, 28 September 2013.

[349] Justin L MacCallum, Alberto Perez, and Ken A Dill. Determining protein structures by combining semireliable data with atomistic physical models by bayesian inference. *Proc. Natl. Acad. Sci. U. S. A.*, 112(22):6985–6990, 2 June 2015.

[350] M Madan Babu. The contribution of intrinsically disordered regions to protein function, cellular complexity, and human disease. *Biochem. Soc. Trans.*, 44(5):1185–1200, 15 October 2016.

[351] Douglas Magde, Elliot Elson, and W W Webb. Thermodynamic fluctuations in a reacting system - measurement by fluorescence correlation spectroscopy. *Phys. Rev. Lett.*, 29(11):705–708, 11 September 1972.

[352] C Magg and F X Schmid. Rapid collapse precedes the fast two-state folding of the cold shock protein. *J. Mol. Biol.*, 335(5):1309–1323, 2004.

[353] Haripada Maity, Mita Maity, Mallela M G Krishna, Leland Mayne, and S Walter Englander. Protein folding: the stepwise assembly of foldon units. *Proc. Natl. Acad. Sci. U. S. A.*, 102(13):4741–4746, 29 March 2005.

[354] Hiranmay Maity and Govardhan Reddy. Folding of protein L with implications for collapse in the denatured state ensemble. *J. Am. Chem. Soc.*, 138(8):2609–2616, 2 March 2016.

[355] G I Makhatadze. Thermodynamics of protein interactions with urea and guanidinium hydrochloride. *J. Phys. Chem. B*, 103(23):4781–4785, 1999.

[356] Liliana Malinovska, Sandra Palm, Kimberley Gibson, Jean-Marc Verbavatz, and Simon Alberti. Dictyostelium discoideum has a highly Q/N-rich proteome and shows an unusual resilience to protein aggregation. *Proc. Natl. Acad. Sci. U. S. A.*, 112(20):E2620–9, 19 May 2015.

[357] M K Malleshaiah, V Shahrezaei, P S Swain, and S W Michnick. The scaffold protein ste5 directly controls a switch-like mating decision in yeast. *Nature*, 465(7294):101–105, 2010.

[358] Matthias Mann and Ole N Jensen. Proteomic analysis of post-translational modifications. *Nat. Biotechnol.*, 21(3):255–261, March 2003.

[359] Albert H Mao, Scott L Crick, Andreas Vitalis, Caitlin L Chicoine, and Rohit V Pappu. Net charge per residue modulates conformational ensembles of intrinsically disordered proteins. *Proceedings of the National Academy of Sciences*, 107(18):8183–8188, 19 April 2010.

[360] Albert H Mao, Nicholas Lyle, and Rohit V Pappu. Describing sequence-ensemble relationships for intrinsically disordered proteins. *Biochem. J*, 449(2):307–318, 15 January 2013.

[361] Albert H Mao and Rohit V Pappu. Crystal lattice properties fully determine short-range interaction parameters for alkali and halide ions. *J. Chem. Phys.*, 137(6):064104, 14 August 2012.

[362] M C Marchetti, J F Joanny, S Ramaswamy, T B Liverpool, J Prost, Madan Rao, and R Aditi Simha. Hydrodynamics of soft active matter. *Rev. Mod. Phys.*, 85(3):1143–1189, 19 July 2013.

[363] J A Marsh, V K Singh, Z Jia, and J D Forman-Kay. Sensitivity of secondary structure propensities to sequence differences between alpha- and gamma-synuclein: implications for fibrillation. *Protein Sci.*, 15(12):2795–2804, 2006.

[364] Joseph A Marsh and Julie D Forman-Kay. Sequence determinants of compaction in intrinsically disordered proteins. *Biophys. J.*, 98(10):2383–2390, 19 May 2010.

[365] M A Martí-Renom, A C Stuart, A Fiser, R Sánchez, F Melo, and A Sali. Comparative protein structure modeling of genes and genomes. *Annu. Rev. Biophys. Biomol. Struct.*, 29:291–325, 2000.

[366] Erik W Martin, Alex S Holehouse, Christy R Grace, Alex Hughes, Rohit V Pappu, and Tanja Mittag. Sequence determinants of the conformational properties of an intrinsically disordered protein prior to and upon multisite phosphorylation. *J. Am. Chem. Soc.*, 138(47):15323–15335, 30 November 2016.

[367] P E Mason, J W Brady, G W Neilson, and C E Dempsey. The interaction of guanidinium ions with a model peptide. *Biophys. J.*, 93(1):L4–L6, 2007.

[368] P E Mason, C E Dempsey, G W Neilson, S R Kline, and J W Brady. Preferential interactions of guanidinum ions with aromatic groups over aliphatic groups. *J. Am. Chem. Soc.*, 131(46):16689–16696, 2009.

[369] Daniel Mateju, Titus M Franzmann, Avinash Patel, Andrii Kopach, Edgar E Boczek, Shovamayee Maharana, Hyun O Lee, Serena Carra, Anthony A Hyman, and Simon Alberti. An aberrant phase transition of stress granules triggered by misfolded protein and prevented by chaperone function. *EMBO J.*, 4 April 2017.

[370] Matthew K Matlock, Alex S Holehouse, and Kristen M Naegle. ProteomeScout: a repository and analysis resource for post-translational modifications and proteins. *Nucleic Acids Res.*, 43(Database issue):D521–30, January 2015.

[371] C R Matthews. Pathways of protein folding. *Annu. Rev. Biochem.*, 62:653–683, 1993.

[372] Sebastian McClendon, Carla C Rospigliosi, and David Eliezer. Charge neutralization and collapse of the c-terminal tail of alpha-synuclein at low ph. *Protein Sci.*, 18(7):1531–1540, July 2009.

[373] E H McConkey. Molecular evolution, intracellular organization, and the quinary structure of proteins. *Proc. Natl. Acad. Sci. U. S. A.*, 79(10):3236–3240, May 1982.

[374] Robert T McGibbon, Kyle A Beauchamp, Matthew P Harrigan, Christoph Klein, Jason M Swails, Carlos X Hernández, Christian R Schwantes, Lee-Ping Wang, Thomas J Lane, and Vijay S Pande. MDTraj: a modern, open library for the analysis of molecular dynamics trajectories. *Biophys. J.*, 109(8):1528–1532, 20 October 2015.

[375] B R McNaughton, J J Cronican, D B Thompson, and D R Liu. Mammalian cell penetration, siRNA transfection, and DNA transfection by supercharged proteins. *Proc. Natl. Acad. Sci. U. S. A.*, 106(15):6111–6116, 2009.

[376] Wenli Meng, Bowu Luan, Nicholas Lyle, Rohit V Pappu, and Daniel P Raleigh. The denatured state ensemble contains significant local and Long-Range structure under native conditions: Analysis of the N-Terminal domain of ribosomal protein L9. *Biochemistry*, 52(15):2662–2671, 16 April 2013.

[377] Wenli Meng, Nicholas Lyle, Bowu Luan, Daniel P Raleigh, and Rohit V Pappu. Experiments and simulations show how long-range contacts can form in expanded unfolded proteins with negligible secondary structure. *Proc. Natl. Acad. Sci. U. S. A.*, 110(6):2123–2128, 5 February 2013.

[378] Davide Mercadante, Sigrid Milles, Gustavo Fuertes, Dmitri I Svergun, Edward A Lemke, and Frauke Gräter. Kirkwood-Buff approach rescues overcollapse of a disordered protein in canonical protein force fields. *J. Phys. Chem. B*, 119(25):7975–7984, 25 June 2015.

[379] Kusai A Merchant, Robert B Best, John M Louis, Irina V Gopich, and William A Eaton. Characterizing the unfolded states of proteins using single-molecule FRET spectroscopy and molecular simulations. *Proceedings of the National Academy of Sciences*, 104(5):1528–1533, 30 January 2007.

[380] Alexandre Mermillod-Blondin, Euan McLeod, and Craig B Arnold. High-speed varifocal imaging with a tunable acoustic gradient index of refraction lens. *Opt. Lett.*, 33(18):2146–2148, 2008.

[381] Lauren Ann Metskas and Elizabeth Rhoades. Conformation and dynamics of the troponin I C-Terminal domain: Combining Single-Molecule and computational approaches for a disordered protein region. *J. Am. Chem. Soc.*, 137(37):11962–11969, 23 September 2015.

[382] Huaiyu Mi, Anushya Muruganujan, and Paul D Thomas. PANTHER in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. *Nucleic Acids Res.*, 41(Database issue):D377–D386, January 2013.

[383] M Miao, C M Bellingham, R J Stahl, E E Sitarz, C J Lane, and F W Keeley. Sequence and structure determinants for the self-aggregation of recombinant polypeptides modeled after human elastin. *J. Biol. Chem.*, 278(49):48553–48562, 2003.

[384] Naveen Michaud-Agrawal, Elizabeth J Denning, Thomas B Woolf, and Oliver Beckstein. MDAnalysis: a toolkit for the analysis of molecular dynamics simulations. *J. Comput. Chem.*, 32(10):2319–2327, 30 July 2011.

[385] Cayla M Miller, Young C Kim, and Jeetain Mittal. Protein composition determines the effect of crowding on the properties of disordered proteins. *Biophys. J.*, 111(1):28–37, 12 July 2016.

[386] Sigrid Milles, Davide Mercadante, Iker Valle Aramburu, Malene Ringkjøbing Jensen, Niccolò Banterle, Christine Koehler, Swati Tyagi, Jane Clarke, Sarah L Shammas, Martin Blackledge, Frauke Gräter, and Edward A Lemke. Plasticity of an ultrafast interaction between nucleoporins and nuclear transport receptors. *Cell*, 163(3):734–745, 22 October 2015.

[387] I S Millett, S Doniach, and K W Plaxco. Toward a taxonomy of the denatured state: small angle scattering studies of unfolded proteins. *Adv. Protein Chem.*, 62:241–262, 2002.

[388] A P Minton. Implications of macromolecular crowding for protein assembly. *Curr. Opin. Struct. Biol.*, 10(1):34–39, February 2000.

[389] Allen P Minton. Models for excluded volume interaction between an unfolded protein and rigid macromolecular cosolutes: Macromolecular crowding and protein stability revisited. *Biophys. J.*, 88(2):971–985, February 2005.

[390] Diana M Mitrea, Jaclyn A Cika, Clifford S Guy, David Ban, Priya R Banerjee, Christopher B Stanley, Amanda Nourse, Ashok A Deniz, and Richard W Kriwacki. Nucleophosmin integrates within the nucleolus via multi-modal interactions with proteins displaying r-rich linear motifs and rRNA. *Elife*, 5, 2 February 2016.

[391] Diana M Mitrea and Richard W Kriwacki. Phase separation in biology; functional organization of a higher order. *Cell Commun. Signal.*, 14(1):1–20, 5 January 2016.

[392] Tanja Mittag, Joseph Marsh, Alexander Grishaev, Stephen Orlicky, Hong Lin, Frank Sicheri, Mike Tyers, and Julie D Forman-Kay. Structure/function implications in a dynamic complex of the intrinsically disordered sic1 with the cdc4 subunit of an SCF ubiquitin ligase. *Structure*, 18(4):494–506, 14 March 2010.

[393] Tanja Mittag, Stephen Orlicky, Wing-Yiu Choy, Xiaojing Tang, Hong Lin, Frank Sicheri, Lewis E Kay, Mike Tyers, and Julie D Forman-Kay. Dynamic equilibrium engagement of a polyvalent ligand with a single-site receptor. *Proceedings of the National Academy of Sciences*, 105(46):17772–17777, 18 November 2008.

[394] Anuradha Mittal, Nicholas Lyle, Tyler S Harmon, and Rohit V Pappu. Hamiltonian switch metropolis monte carlo simulations for improved conformational sampling of intrinsically disordered regions tethered to ordered domains of proteins. *J. Chem. Theory Comput.*, 10(8):3550–3562, 12 August 2014.

[395] S J Miyake-Stoner, A M Miller, J T Hammill, J C Peeler, K R Hess, R A Mehl, and S H Brewer. Probing protein folding using site-specifically encoded unnatural amino acids as FRET donors with tryptophan. *Biochemistry*, 48(25):5953–5962, 2009.

[396] S Miyamoto and P A Kollman. Settle - an analytical version of the shake and rattle algorithm for rigid water models. *J. Comput. Chem.*, 13(8):952–962, 1992.

[397] Beate Moeser and Dominik Horinek. Unified description of urea denaturation: Backbone and side chains contribute equally in the transfer model. *J. Phys. Chem. B*, 118(1):107–114, 9 January 2014.

[398] Y K Mok, C M Kay, L E Kay, and J Forman-Kay. NOE data demonstrating a compact unfolded state for an SH3 domain under non-denaturing conditions. *J. Mol. Biol.*, 289(3):619–638, 11 June 1999.

[399] Amandine Molliex, Jamshid Temirov, Jihun Lee, Maura Coughlin, Anderson P Kanagaraj, Hong Joo Kim, Tanja Mittag, and J Paul Taylor. Phase separation by low complexity domains promotes stress granule assembly and drives pathological fibrillization. *Cell*, 163(1):123–133, 24 September 2015.

[400] Kevin A Morano, Chris M Grant, and W Scott Moye-Rowley. The response to heat shock and oxidative stress in saccharomyces cerevisiae. *Genetics*, 190(4):1157–1195, April 2012.

[401] Joseph A Morrone, Alberto Perez, Justin MacCallum, and Ken A Dill. Computed binding of peptides to proteins with MELD-Accelerated molecular dynamics. *J. Chem. Theory Comput.*, 19 January 2017.

[402] Alexandra Moura, Michael A Savageau, and Rui Alves. Relative amino acid composition signatures of organisms and environments. *PLoS One*, 8(10):e77319, 25 October 2013.

[403] Debashish Mukherji, Carlos M Marques, and Kurt Kremer. Polymer collapse in miscible good solvents is a generic phenomenon driven by preferential adsorption. *Nat. Commun.*, 5:4882, 12 September 2014.

[404] Samrat Mukhopadhyay, Rajaraman Krishnan, Edward A Lemke, Susan Lindquist, and Ashok A Deniz. A natively unfolded yeast prion monomer adopts an ensemble of collapsed and rapidly fluctuating structures. *Proc. Natl. Acad. Sci. U. S. A.*, 104(8):2649–2654, 20 February 2007.

[405] S Muller-Spath, A Soranno, V Hirschfeld, H Hofmann, S Ruegger, L Reymond, D Nettels, and B Schuler. Charge interactions can dominate the dimensions of intrinsically disordered proteins. *Proc. Natl. Acad. Sci. U. S. A.*, 107(33):14609–14614, 2010.

[406] Tetsuro Murakami, Seema Qamar, Julie Qiaojin Lin, Gabriele S Kaminski Schierle, Eric Rees, Akinori Miyashita, Ana R Costa, Roger B Dodd, Fiona T S Chan, Claire H Michel, Deborah Kronenberg-Versteeg, Yi Li, Seung-Pil Yang, Yosuke Wakutani, William Meadows, Rodylyn Rose Ferry, Liang Dong, Gian Gaetano Tartaglia, Giorgio Favrin, Wen-Lang Lin, Dennis W Dickson, Mei Zhen, David Ron, Gerold Schmitt-Ulms, Paul E Fraser, Neil A Shneider, Christine Holt, Michele Vendruscolo, Clemens F Kaminski, and Peter St George-Hyslop. ALS/FTD mutation-induced phase transition of FUS liquid droplets and reversible hydrogels into irreversible hydrogels impairs RNP granule function. *Neuron*, 88(4):678–690, 2015.

[407] L R Murphy, A Wallqvist, and R M Levy. Simplified amino acid alphabets for protein fold recognition and implications for folding. *Protein Eng.*, 13(3):149–152, March 2000.

[408] A G Murzin, S E Brenner, T Hubbard, and C Chothia. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, 247(4):536–540, 7 April 1995.

[409] M Muthukumar. Thermodynamics of polymer solutions. *J. Chem. Phys.*, 85(8):4722–4728, 15 October 1986.

[410] M Muthukumar and S F Edwards. Extrapolation formulas for polymer solution properties. *J. Chem. Phys.*, 76(5):2720–2730, 1 March 1982.

[411] Efstratios Mylonas, Antje Hascher, Pau Bernadó, Martin Blackledge, Eckhard Mandelkow, and Dmitri I Svergun. Domain conformation of tau protein studied by solution small-angle x-ray scattering. *Biochemistry*, 47(39):10345–10353, 30 September 2008.

[412] Arjun Narayanan, Anatoli B Meriin, Michael Y Sherman, and Ibrahim I Cisse. A first order phase transition underlies the formation of Sub-Diffractive protein aggregates in mammalian cells. 19 June 2017.

[413] A H Narten. Liquid water: Atom pair correlation functions from neutron and XRay diffraction. *J. Chem. Phys.*, 56(11):5681–5687, 1 June 1972.

[414] Eviatar Natan, Jonathan N Wells, Sarah A Teichmann, and Joseph A Marsh. Regulation, evolution and consequences of cotranslational protein complex assembly. *Curr. Opin. Struct. Biol.*, 42:90–97, February 2017.

[415] S B Needleman and C D Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.*, 48(3):443–453, March 1970.

[416] Pauline C Ng and Steven Henikoff. SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res.*, 31(13):3812–3814, 1 July 2003.

[417] Katsuyoshi Nishinari. Some thoughts on the definition of a gel. In *Gels: Structures, Properties, and Functions*, pages 87–94. Springer, Berlin, Heidelberg, 2009.

[418] Izumi Nishio, Shao-Tang Sun, Gerald Swislow, and Toyoichi Tanaka. First observation of the coil–globule transition in a single polymer chain. *Nature*, 281(5728):208–209, 20 September 1979.

[419] Daniel A Nissley, Ajeet K Sharma, Nabeel Ahmed, Ulrike A Friedrich, Günter Kramer, Bernd Bukau, and Edward P O'Brien. Accurate prediction of cellular co-translational folding indicates proteins can switch from post- to co-translational folding. *Nat. Commun.*, 7:10341, 18 February 2016.

[420] Timothy J Nott, Timothy D Craggs, and Andrew J Baldwin. Membraneless organelles can melt nucleic acid duplexes and act as biomolecular filters. *Nat. Chem.*, 8(6):569–575, June 2016.

[421] Timothy J Nott, Evangelia Petsalaki, Patrick Farber, Dylan Jervis, Eden Fussner, Anne Plochowietz, Timothy D Craggs, David P Bazett-Jones, Tony Pawson, Julie D Forman-Kay, and Andrew J Baldwin. Phase transition of a disordered nuage protein generates environmentally responsive membraneless organelles. *Mol. Cell*, 57(5):936–947, 5 March 2015.

[422] D Novick, B Cohen, and M Rubinstein. The human interferon alpha/beta receptor: characterization and molecular cloning. *Cell*, 77(3):391–400, 6 May 1994.

[423] Nathaniel V Nucci, Maxim S Pometun, and A Joshua Wand. Mapping the hydration dynamics of ubiquitin. *J. Am. Chem. Soc.*, 133(32):12326–12329, 17 August 2011.

[424] Nathaniel V Nucci, Maxim S Pometun, and A Joshua Wand. Site-resolved measurement of water-protein interactions by solution NMR. *Nat. Struct. Mol. Biol.*, 18(2):245–249, February 2011.

[425] Bo Nyström, Anna-Lena Kjøniksen, Neda Beheshti, Atoosa Maleki, Kaizheng Zhu, Kenneth D Knudsen, Ramón Pamies, José G Hernández Cifre, and José García de la Torre. Characterization of polyelectrolyte features in polysaccharide systems and mucin. *Adv. Colloid Interface Sci.*, 158(1-2):108–118, 12 July 2010.

[426] Oates, M.E., Romero, P., Ishida, T., Ghalwash, M., Mizianty, M.J., Xue, B., Dosztányi, S., Uversky, V.N., Obradovic, Z., Kurgan, L., Dunker, A.K., Gough, and J. D2P2: Database of disordered protein predictions. *Nucleic Acids Res.*, 41:D508–D516, 2013.

[427] E P O'Brien, B R Brooks, and D Thirumalai. Molecular origin of constant m-values, denatured state collapse, and Residue-Dependent transition midpoints in globular proteins. *Biochemistry*, 48(17):3743–3754, 2009.

[428] E P O'Brien, G Morrison, B R Brooks, and D Thirumalai. How accurate are polymer models in the analysis of forster resonance energy transfer experiments on proteins? *J. Chem. Phys.*, 130(12):124903, 2009.

[429] Edward P O'Brien, Ruxandra I Dima, Bernard Brooks, and D Thirumalai. Interactions between hydrophobic and ionic solutes in aqueous guanidinium chloride and urea solutions: lessons for protein denaturation mechanism. *J. Am. Chem. Soc.*, 129(23):7346–7353, 13 June 2007.

[430] Edward P O'Brien, Guy Ziv, Gilad Haran, Bernard R Brooks, and D Thirumalai. Effects of denaturants and osmolytes on proteins are accurately predicted by the molecular transfer model. *Proc. Natl. Acad. Sci. U. S. A.*, 105(36):13403–13408, 29 August 2008.

[431] José Nelson Onuchic and Peter G Wolynes. Theory of protein folding. *Curr. Opin. Struct. Biol.*, 14(1):70–75, February 2004.

[432] E K O'Shea, J D Klemm, P S Kim, and T Alber. X-ray structure of the GCN4 leucine zipper, a two-stranded, parallel coiled coil. *Science*, 254(5031):539–544, 25 October 1991.

[433] Sergey Ovchinnikov, Hahnbeom Park, Neha Varghese, Po-Ssu Huang, Georgios A Pavlopoulos, David E Kim, Hetunandan Kamisetty, Nikos C Kyrpides, and David Baker. Protein structure determination using metagenome sequence data. *Science*, 355(6322):294–298, 20 January 2017.

[434] Valéry Ozenne, Frédéric Bauer, Loïc Salmon, Jie-Rong Huang, Malene Ringkjøbing Jensen, Stéphane Segard, Pau Bernadó, Céline Charavay, and Martin Blackledge. Flexible-meccano: a tool for the generation of explicit ensemble descriptions of intrinsically disordered proteins and their associated experimental observables. *Bioinformatics*, 28(11):1463–1470, 1 June 2012.

[435] C N Pace. Contribution of the hydrophobic effect to globular protein stability. *J. Mol. Biol.*, 226(1):29–35, 5 July 1992.

[436] Chi W Pak, Martyna Kosno, Alex S Holehouse, Shae B Padrick, Anuradha Mittal, Rustam Ali, Ali A Yunus, David R Liu, Rohit V Pappu, and Michael K Rosen. Sequence determinants of intracellular phase separation by complex coacervation of a disordered protein. *Mol. Cell*, 63(1):72–85, 7 July 2016.

[437] Rita Pancsa and Peter Tompa. Structural disorder in eukaryotes. *PLoS One*, 7(4):e34687, 5 April 2012.

[438] Vijay S Pande and Daniel S Rokhsar. Molecular dynamics simulations of unfolding and refolding of a β-hairpin fragment of protein G. *Proceedings of the National Academy of Sciences*, 96(16):9062–9067, 3 August 1999.

[439] H C Pant, Veeranna, and P Grant. Regulation of axonal neurofilament phosphorylation. *Curr. Top. Cell. Regul.*, 36:133–150, 2000.

[440] Rohit V. Pappu, Xiaoling Wang, Andreas Vitalis, and Scott L. Crick. A polymer physics perspective on driving forces and mechanisms for protein aggregation - highlight issue: Protein folding. *Arch. Biochem. Biophys.*, 469(1):132–141, January 2008.

[441] Giacomo Parigi, Nasrollah Rezaei-Ghaleh, Andrea Giachetti, Stefan Becker, Claudio Fernandez, Martin Blackledge, Christian Griesinger, Markus Zweckstetter, and Claudio Luchinat. Long-range correlated dynamics in intrinsically disordered proteins. *J. Am. Chem. Soc.*, 136(46):16201–16209, 19 November 2014.

[442] M Parrinello and A Rahman. Polymorphic transitions in Single-Crystals - a new Molecular-Dynamics method. *J. Appl. Phys.*, 52(12):7182–7190, 1981.

[443] Avinash Patel, Hyun O Lee, Louise Jawerth, Shovamayee Maharana, Marcus Jahnel, Marco Y Hein, Stoyno Stoynov, Julia Mahamid, Shambaditya Saha, Titus M Franzmann, Andrej Pozniakovski, Ina Poser, Nicola Maghelli, Loic A Royer, Martin Weigert, Eugene W Myers, Stephan Grill, David Drechsel, Anthony A Hyman, and Simon Alberti. A liquid-to-solid phase transition of the ALS protein FUS accelerated by disease mutation. *Cell*, 162(5):1066–1077, 27 August 2015.

[444] Avinash Patel, Liliana Malinovska, Shambaditya Saha, Jie Wang, Simon Alberti, Yamuna Krishnan, and Anthony A Hyman. ATP as a biological hydrotrope. *Science*, 356(6339):753–756, 19 May 2017.

[445] Alexandru Patriciu, Gregory S Chirikjian, and Rohit V Pappu. Analysis of the conformational dependence of mass-metric tensor determinants in serial polymers with constraints. *J. Chem. Phys.*, 121(24):12708–12720, 22 December 2004.

[446] Dennis Perchak, J Skolnick, and Robert Yaris. Dynamics of rigid and flexible constraints for polymers. effect of the fixman potential. *Macromolecules*, 18(3):519–525, 1 March 1985.

[447] Alberto Perez, Joseph A Morrone, Carlos Simmerling, and Ken A Dill. Advances in free-energy-based simulations of protein folding and ligand binding. *Curr. Opin. Struct. Biol.*, 36:25–31, February 2016.

[448] Juan R Perilla, Boon Chong Goh, C Keith Cassidy, Bo Liu, Rafael C Bernardi, Till Rudack, Hang Yu, Zhe Wu, and Klaus Schulten. Molecular dynamics simulations of large macromolecular complexes. *Curr. Opin. Struct. Biol.*, 31:64–74, April 2015.

[449] Hemali P Phatnani and Arno L Greenleaf. Phosphorylation and functions of the RNA polymerase II CTD. *Genes Dev.*, 20(21):2922–2936, 1 November 2006.

[450] Patrick C Phillips. Epistasis—the essential role of gene interactions in the structure and evolution of genetic systems. *Nat. Rev. Genet.*, 9(11):855–867, 2008.

[451] R. Phillips, J. Kondev, J. Theriot, and N. Orme. *Physical Biology of the Cell*. Garland Science, 2013.

[452] Stefano Piana, Alexander G Donchev, Paul Robustelli, and David E Shaw. Water dispersion interactions strongly influence simulated structural properties of disordered protein states. *J. Phys. Chem. B*, 119(16):5113–5123, 2015.

[453] Stefano Piana, John L Klepeis, and David E Shaw. Assessing the accuracy of physical models used in protein-folding simulations: quantitative evidence from long molecular dynamics simulations. *Curr. Opin. Struct. Biol.*, 24:98–105, February 2014.

[454] V Pierce, M Kang, M Aburi, S Weerasinghe, and P E Smith. Recent applications of Kirkwood-Buff theory to biological systems. *Cell Biochem. Biophys.*, 50(1):1–22, 2008.

[455] Damiano Piovesan, Francesco Tabaro, Ivan Mičetić, Marco Necci, Federica Quaglia, Christopher J Oldfield, Maria Cristina Aspromonte, Norman E Davey, Radoslav Davidović, Zsuzsanna Dosztányi, Arne Elofsson, Alessandra Gasparini, András Hatos, Andrey V Kajava, Lajos Kalmar, Emanuela Leonardi, Tamas Lazar, Sandra Macedo-Ribeiro, Mauricio Macossay-Castillo, Attila Meszaros, Giovanni Minervini, Nikoletta Murvai, Jordi Pujols, Daniel B Roche, Edoardo Salladini, Eva Schad, Antoine Schramm, Beata Szabo, Agnes Tantos, Fiorella Tonello, Konstantinos D Tsirigos, Nevena Veljković, Salvador Ventura, Wim Vranken, Per Warholm, Vladimir N Uversky, A Keith Dunker, Sonia Longhi, Peter Tompa, and Silvio C E Tosatto. DisProt 7.0: a major update of the database of disordered proteins. *Nucleic Acids Res.*, 45(D1):D1123–D1124, 4 January 2017.

[456] M Pitschke, R Prior, M Haupt, and D Riesner. Detection of single amyloid beta-protein aggregates in the cerebrospinal fluid of alzheimer's patients by fluorescence correlation spectroscopy. *Nat. Med.*, 4(7):832–834, 1998.

[457] K W Plaxco, I S Millett, D J Segel, S Doniach, and D Baker. Chain collapse can occur concomitantly with the rate-limiting step in protein folding. *Nat. Struct. Biol.*, 6(6):554–556, June 1999.

[458] S S Plotkin. Speeding protein folding beyond the g(o) model: how a little frustration sometimes helps. *Proteins*, 45(4):337–345, 2001.

[459] M H Polymeropoulos, C Lavedan, E Leroy, S E Ide, A Dehejia, A Dutra, B Pike, H Root, J Rubenstein, R Boyer, E S Stenroos, S Chandrasekharappa, A Athanassiadou, T Papapetropoulos, W G Johnson, A M Lazzarini, R C Duvoisin, G Di Iorio, L I Golbe, and R L Nussbaum. Mutation in the alpha-synuclein gene identified in families with parkinson's disease. *Science*, 276(5321):2045–2047, 27 June 1997.

[460] Bede Portz, Feiyue Lu, Eric B Gibbs, Joshua E Mayfield, M Rachel Mehaffey, Yan Jessie Zhang, Jennifer S Brodbelt, Scott A Showalter, and David S Gilmour. Structural heterogeneity in the intrinsically disordered RNA polymerase II c-terminal domain. *Nat. Commun.*, 8:15231, 12 May 2017.

[461] Emilio Potenza, Tomás Di Domenico, Ian Walsh, and Silvio C E Tosatto. MobiDB 2.0: an improved database of intrinsically disordered and mobile proteins. *Nucleic Acids Res.*, 43(Database issue):D315–D320, January 2015.

[462] Sander Pronk, Szilard Pall, Roland Schulz, Per Larsson, Par Bjelkmar, Rossen Apostolov, Michael R Shirts, Jeremy C Smith, Peter M Kasson, David van der Spoel, Berk Hess, and Erik Lindahl. GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics*, 29(7):845–854, 2013.

[463] David S W Protter, Bhalchandra S Rao, Briana Van Treeck, Yuan Lin, Laura Mizoue, Michael K Rosen, and Roy Parker. Intrinsically disordered regions contribute promiscuous interactions to RNP granule assembly. 13 June 2017.

[464] S W Provencher and J Glöckner. Estimation of globular protein secondary structure from circular dichroism. *Biochemistry*, 20(1):33–37, 6 January 1981.

[465] O B Ptitsyn. Stages in the mechanism of self-organization of protein molecules. *Dokl. Akad. Nauk SSSR*, 210(5):1213–1215, 11 June 1973.

[466] O B Ptitsyn, R H Pain, G V Semisotnov, E Zerovnik, and O I Razgulyaev. Evidence for a molten globule state as a general intermediate in protein folding. *FEBS Lett.*, 262(1):20–24, 12 March 1990.

[467] Christopher D Putnam, Michal Hammel, Greg L Hura, and John A Tainer. X-ray solution scattering (SAXS) combined with crystallography and computation: defining accurate macromolecular structures, conformations and assemblies in solution. *Q. Rev. Biophys.*, 40(03):191–285, 2007.

[468] Phoebe X Qi, Tobin R Sosnick, and S Walter Englander. The burst phase in ribonuclease a folding and solvent dependence of the unfolded state. *Nat. Struct. Biol.*, 5(10):882–884, October 1998.

[469] Felipe García Quiroz and Ashutosh Chilkoti. Sequence heuristics to encode phase behaviour in intrinsically disordered protein polymers. *Nat. Mater.*, 14(11):1164–1171, November 2015.

[470] Joshua A Rackers, Qiantao Wang, Chengwen Liu, Jean-Philip Piquemal, Pengyu Ren, and Jay W Ponder. An optimized charge penetration model for use with the AMOEBA force field. *Phys. Chem. Chem. Phys.*, 19(1):276–291, 21 December 2016.

[471] Aditya Radhakrishnan, Andreas Vitalis, Albert H Mao, Adam T Steffen, and Rohit V Pappu. Improved atomistic monte carlo simulations demonstrate that poly-l-proline adopts heterogeneous ensembles of conformations of semi-rigid segments interrupted by kinks. *J. Phys. Chem. B*, 116(23):6862–6871, 14 June 2012.

[472] G Raos and G Allegra. Chain collapse and phase separation in poor-solvent polymer solutions: A unified molecular description. *J. Chem. Phys.*, 104(4):1626–1645, 1996.

[473] G Raos and G Allegra. Chain interactions in poor-solvent polymer solutions: Equilibrium and nonequilibrium aspects. *Macromolecules*, 29(20):6663–6670, 1996.

[474] Alpan Raval, Stefano Piana, Michael P Eastwood, and David E Shaw. Assessment of the utility of contact-based restraints in accelerating the prediction of protein structure using molecular dynamics simulations. *Protein Sci.*, 25(1):19–29, 2016.

[475] V Receveur-Brechot and D Durand. How random are intrinsically disordered proteins? a small angle scattering perspective. *Curr. Protein Pept. Sci.*, 13(1):55–75, 2012.

[476] M T Record, E Guinn, L Pegram, and M Capp. Introductory lecture: Interpreting and predicting hofmeister salt ion and solute effects on biopolymer and model processes using the solute partitioning model. *Faraday Discuss.*, 160:9–44, 2013.

[477] Govardhan Reddy and Dave Thirumalai. Collapse precedes folding in Denaturant-Dependent assembly of ubiquitin. *J. Phys. Chem. B*, 11 January 2017.

[478] H Reiersen and A R Rees. The hunchback and its neighbours: proline as an environmental modulator. *Trends Biochem. Sci.*, 26(11):679–684, 2001.

[479] E A Reits and J J Neefjes. From fixed to FRAP: measuring protein mobility and activity in living cells. *Nat. Cell Biol.*, 3(6):E145–7, June 2001.

[480] T L Religa, J S Markson, U Mayor, S M V Freund, and A R Fersht. Solution structure of a protein denatured state and folding intermediate. *Nature*, 437(7061):1053–1056, 13 October 2005.

[481] Y L A Rezus and H J Bakker. Effect of urea on the structural dynamics of water. *Proc. Natl. Acad. Sci. U. S. A.*, 103(49):18417–18420, 20 November 2006.

[482] Gale Rhodes. *Crystallography Made Crystal Clear: A Guide for Users of Macromolecular Models*. Academic Press, San Diego; London, 1993.

[483] Joshua A Riback, Christopher D Katanski, Jamie L Kear-Scott, Evgeny V Pilipenko, Alexandra E Rojek, Tobin R Sosnick, and D Allan Drummond. Stress-Triggered phase separation is an adaptive, evolutionarily tuned response. *Cell*, 168(6):1028–1040.e19, 9 March 2017.

[484] F M Richards. Areas, volumes, packing and protein structure. *Annu. Rev. Biophys. Bioeng.*, 6:151–176, 1977.

[485] Frederic M Richards and Craig E Kundrot. Identification of structural motifs from protein coordinate data: secondary structure and first-level supersecondary structure. *Proteins: Struct. Funct. Bioinf.*, 3(2):71–84, 1988.

[486] J S Richardson. Describing patterns of protein tertiary structure. *Methods Enzymol.*, 115:341–358, 1985.

[487] Jonas Ries and Petra Schwille. New concepts for fluorescence correlation spectroscopy on membranes. *Phys. Chem. Chem. Phys.*, 10(24):3487–3497, 2008.

[488] G Rigaut, A Shevchenko, B Rutz, M Wilm, M Mann, and B Séraphin. A generic protein purification method for protein complex characterization and proteome exploration. *Nat. Biotechnol.*, 17(10):1030–1032, October 1999.

[489] Stefan Roberts, Michael Dzuricky, and Ashutosh Chilkoti. Elastin-like polypeptides as models of intrinsically disordered proteins. *FEBS Lett.*, 589(19 Pt A):2477–2486, 14 September 2015.

[490] Paul Robustelli, Stefano Piana, and David E Shaw. Developing force fields for the accurate simulation of both ordered and disordered protein states. *Biophys. J.*, 112(3):175a, 3 February 2017.

[491] H Roder, G A Elöve, and S W Englander. Structural characterization of folding intermediates in cytochrome c by h-exchange labelling and proton NMR. *Nature*, 335(6192):700–704, 20 October 1988.

[492] H Roder, K Maki, and H Cheng. Early events in protein folding explored by rapid mixing methods. *Chem. Rev.*, 106(5):1836–1861, 2006.

[493] Ofer Rog, Simone Köhler, and Abby F Dernburg. The synaptonemal complex has liquid crystalline properties and spatially regulates meiotic recombination factors. *Elife*, 6, 3 January 2017.

[494] Joseph M Rogers, Annette Steward, and Jane Clarke. Folding and binding of an intrinsically disordered protein: fast, but not 'diffusion-limited'. *J. Am. Chem. Soc.*, 135(4):1415–1422, 30 January 2013.

[495] Carol A Rohl, Charlie E M Strauss, Kira M S Misura, and David Baker. Protein structure prediction using rosetta. *Methods Enzymol.*, 383:66–93, 2004.

[496] Geoffrey C Rollins and Ken A Dill. General mechanism of two-state protein folding kinetics. *J. Am. Chem. Soc.*, 136(32):11420–11427, 13 August 2014.

[497] G D Rose, L M Gierasch, and J A Smith. Turns in peptides and proteins. *Adv. Protein Chem.*, 37:1–109, 1985.

[498] M N Rosenbluth, A H Teller, and E Teller. Equation of state calculations by fast computing machines. *J. Chem. Phys.*, 21(6):1087–1092, 1 June 1953.

[499] J Rösgen, B M Pettitt, and D W Bolen. Uncovering the basis for nonideal behavior of biological molecules. *Biochemistry*, 43(45):14472–14484, 2004.

[500] B Rost. Review: protein secondary structure prediction continues to rise. *J. Struct. Biol.*, 134(2-3):204–218, May 2001.

[501] Edina Rosta and Gerhard Hummer. Error and efficiency of replica exchange molecular dynamics simulations. *J. Chem. Phys.*, 131(16):165102, 28 October 2009.

[502] Edina Rosta and Gerhard Hummer. Error and efficiency of simulated tempering simulations. *J. Chem. Phys.*, 132(3):034102, 21 January 2010.

[503] Prince E Rouse. A theory of the linear viscoelastic properties of dilute solutions of coiling polymers. *J. Chem. Phys.*, 21(7):1272–1280, 1 July 1953.

[504] M Rubinstein and Ralph H Colby. *Polymer Physics*. Oxford University PRess, New York, 2003.

[505] Michael Rubinstein and Alexander N Semenov. Thermoreversible gelation in solutions of associating polymers. 2. linear dynamics. *Macromolecules*, 31(4):1386–1397, 1 February 1998.

[506] Kiersten M Ruff, Tyler S Harmon, and Rohit V Pappu. CAMELOT: A machine learning approach for coarse-grained simulations of aggregation of block-copolymeric protein sequences. *J. Chem. Phys.*, 143(24):243123, 28 December 2015.

[507] Kiersten M Ruff and Alex S Holehouse. SAXS versus FRET: A matter of heterogeneity? *Biophys. J.*, 15 August 2017.

[508] Kiersten M Ruff, Siddique J Khan, and Rohit V Pappu. A coarse-grained model for polyglutamine aggregation modulated by amphipathic flanking sequences. *Biophys. J.*, 107(5):1226–1235, 2 September 2014.

[509] M Sadqi, L J Lapidus, and V Munoz. How fast is protein hydrophobic collapse? *Proc. Natl. Acad. Sci. U. S. A.*, 100(21):12117–12122, 2003.

[510] Shambaditya Saha, Christoph A Weber, Marco Nousch, Omar Adame-Arana, Carsten Hoege, Marco Y Hein, Erin Osborne-Nishimura, Julia Mahamid, Marcus Jahnel, Louise Jawerth, Andrej Pozniakovski, Christian R Eckmann, Frank Jülicher, and Anthony A Hyman. Polar positioning of phase-separated liquid compartments in cells regulated by an mRNA competition mechanism. *Cell*, 166(6):1572–1584.e16, 8 September 2016.

[511] Helen Saibil. Chaperone machines for protein folding, unfolding and disaggregation. *Nat. Rev. Mol. Cell Biol.*, 14(10):630–642, 12 September 2013.

[512] Daisuke Sakakibara, Atsuko Sasaki, Teppei Ikeya, Junpei Hamatsu, Tomomi Hanashima, Masaki Mishima, Masatoshi Yoshimasu, Nobuhiro Hayashi, Tsutomu Mikawa, Markus Wälchli, Brian O Smith, Masahiro Shirakawa, Peter Güntert, and Yutaka Ito. Protein structure determination in living cells by in-cell NMR spectroscopy. *Nature*, 458(7234):102–105, 5 March 2009.

[513] Loïc Salmon, Gabrielle Nodet, Valéry Ozenne, Guowei Yin, Malene Ringkjøbing Jensen, Markus Zweckstetter, and Martin Blackledge. NMR characterization of long-range order in intrinsically disordered proteins. *J. Am. Chem. Soc.*, 132(24):8407–8418, 23 June 2010.

[514] Romelia Salomon-Ferrer, Andreas W Götz, Duncan Poole, Scott Le Grand, and Ross C Walker. Routine microsecond molecular dynamics simulations with AMBER on GPUs. 2. explicit solvent particle mesh ewald. *J. Chem. Theory Comput.*, 9(9):3878–3888, 10 September 2013.

[515] David W Sanders, Sarah K Kaufman, Sarah L DeVos, Apurwa M Sharma, Hilda Mirbaha, Aimin Li, Scarlett J Barker, Alex C Foley, Julian R Thorpe, Louise C Serpell, Timothy M Miller, Lea T Grinberg, William W Seeley, and Marc I Diamond. Distinct tau prion strains propagate in cells and mice and define different tauopathies. *Neuron*, 82(6):1271–1288, 18 June 2014.

[516] David F Savage, Bruno Afonso, Anna H Chen, and Pamela A Silver. Spatially ordered dynamics of the bacterial carbon fixation machinery. *Science*, 327(5970):1258–1261, 5 March 2010.

[517] Lucas Sawle and Kingshuk Ghosh. A theoretical method to compute sequence dependent configurational properties in charged polymers and proteins. *J. Chem. Phys.*, 143(8):085101, 28 August 2015.

[518] B Schobert and H Tschesche. Unusual solution properties of proline and its interaction with proteins. *Biochim. Biophys. Acta*, 541(2):270–277, 15 June 1978.

[519] Friedrich Schotte, Jayashree Soman, John S Olson, Michael Wulff, and Philip A Anfinrud. Picosecond time-resolved x-ray crystallography: probing protein function in real time. *J. Struct. Biol.*, 147(3):235–246, September 2004.

[520] B Schuler, E A Lipman, P J Steinbach, M Kumke, and W A Eaton. Polyproline and the "spectroscopic ruler" revisited with single-molecule fluorescence. *Proc. Natl. Acad. Sci. U. S. A.*, 102(8):2754–2759, 2005.

[521] Benjamin Schuler and Hagen Hofmann. Single-molecule spectroscopy of protein folding dynamics—expanding scope and timescales - folding and binding / protein-nucleic acid interactions. *Curr. Opin. Struct. Biol.*, 23(1):36–47, February 2013.

[522] Benjamin Schuler, Everett A Lipman, and William A Eaton. Probing the free-energy surface for protein folding with single-molecule fluorescence spectroscopy. *Nature*, 419(6908):743–747, 17 October 2002.

[523] Benjamin Schuler, Andrea Soranno, Hagen Hofmann, and Daniel Nettels. Single-Molecule FRET spectroscopy and the polymer physics of unfolded and intrinsically disordered proteins. *Annu. Rev. Biophys.*, 45:207–231, 5 July 2016.

[524] B A Schulman, P S Kim, C M Dobson, and C Redfield. A residue-specific NMR view of the non-cooperative unfolding of a molten globule. *Nat. Struct. Biol.*, 4(8):630–634, August 1997.

[525] Martin Schwalbe, Harindranath Kadavath, Jacek Biernat, Valery Ozenne, Martin Blackledge, Eckhard Mandelkow, and Markus Zweckstetter. Structural impact of tau phosphorylation at threonine 231. *Structure*, 23(8):1448–1458, 4 August 2015.

[526] Martin Schwalbe, Valéry Ozenne, Stefan Bibow, Mariusz Jaremko, Lukasz Jaremko, Michal Gajda, Malene Ringkjøbing Jensen, Jacek Biernat, Stefan Becker, Eckhard Mandelkow, Markus Zweckstetter, and Martin Blackledge. Predictive atomic resolution descriptions of intrinsically disordered htau40 and α-synuclein in solution from NMR and small angle scattering. *Structure*, 22(2):238–249, 4 February 2014.

[527] Jacob C Schwartz, Christopher C Ebmeier, Elaine R Podell, Joseph Heimiller, Dylan J Taatjes, and Thomas R Cech. FUS binds the CTD of RNA polymerase II and regulates its phosphorylation at ser2. *Genes Dev.*, 26(24):2690–2695, 15 December 2012.

[528] Jacob C Schwartz, Xueyin Wang, Elaine R Podell, and Thomas R Cech. RNA seeds higher-order assembly of FUS protein. *Cell Rep.*, 5(4):918–925, 27 November 2013.

[529] Thomas Schwarz-Romond, Christien Merrifield, Benjamin J Nichols, and Mariann Bienz. The wnt signalling effector dishevelled forms dynamic protein assemblies rather than stable associations with cytoplasmic vesicles. *J. Cell Sci.*, 118(Pt 22):5269–5277, 15 November 2005.

[530] Charles D Schwieters, John J Kuszewski, Nico Tjandra, and G Marius Clore. The Xplor-NIH NMR molecular structure determination package. *J. Magn. Reson.*, 160(1):65–73, January 2003.

[531] E Schwob, T Bohm, M D Mendenhall, and K Nasmyth. The b-type cyclin kinase inhibitor p40SIC1 controls the G1 to S transition in s. cerevisiae. *Cell*, 79(2):233–244, 1994.

[532] Alexander N Semenov* and Michael Rubinstein. Thermoreversible gelation in solutions of associative polymers. 1. statics. *Macromolecules*, 31(4):1373–1385, 1998.

[533] Maya Shamir, Yinon Bar-On, Rob Phillips, and Ron Milo. SnapShot: Timescales in cell biology. *Cell*, 164(6):1302–1302.e1, 10 March 2016.

[534] Sarah L Shammas, Michael D Crabtree, Liza Dahal, Basile I M Wicky, and Jane Clarke. Insights into coupled folding and binding mechanisms from kinetic studies. *J. Biol. Chem.*, 291(13):6689–6695, 2016.

[535] David E Shaw, Martin M Deneroff, Ron O Dror, Jeffrey S Kuskin, Richard H Larson, John K Salmon, Cliff Young, Brannon Batson, Kevin J Bowers, Jack C Chao, Michael P Eastwood, Joseph Gagliardo, J P Grossman, C Richard Ho, Douglas J Ierardi, István Kolossváry, John L Klepeis, Timothy Layman, Christine McLeavey, Mark A Moraes, Rolf Mueller, Edward C Priest, Yibing Shan, Jochen Spengler, Michael Theobald, Brian Towles, and Stanley C Wang. Anton, a special-purpose machine for molecular dynamics simulation. *Commun. ACM*, 51(7):91–97, July 2008.

[536] David E Shaw, Paul Maragakis, Kresten Lindorff-Larsen, Stefano Piana, Ron O Dror, Michael P Eastwood, Joseph A Bank, John M Jumper, John K Salmon, Yibing Shan, and Willy Wriggers. Atomic-level characterization of the structural dynamics of proteins. *Science*, 330(6002):341–346, 15 October 2010.

[537] Felix B Sheinerman, Raquel Norel, and Barry Honig. Electrostatic aspects of protein–protein interactions. *Curr. Opin. Struct. Biol.*, 10(2):153–159, 1 April 2000.

[538] Eilon Sherman and Gilad Haran. Coil-globule transition in the denatured state of a small protein. *Proceedings of the National Academy of Sciences*, 103(31):11539–11543, August 2006.

[539] Eilon Sherman, Anna Itkin, Yosef Yehuda Kuttner, Elizabeth Rhoades, Dan Amir, Elisha Haas, and Gilad Haran. Using fluorescence correlation spectroscopy to study conformational changes in denatured proteins. *Biophys. J.*, 94(12):4819–4827, 7 March 2008.

[540] Yue Shi, Zhen Xia, Jiajing Zhang, Robert Best, Chuanjie Wu, Jay W Ponder, and Pengyu Ren. The polarizable atomic multipole-based AMOEBA force field for proteins. *J. Chem. Theory Comput.*, 9(9):4046–4063, 2013.

[541] Zhengshuang Shi, Kang Chen, Zhigang Liu, and Neville R Kallenbach. Conformation of the backbone in unfolded proteins. *Chem. Rev.*, 106(5):1877–1897, May 2006.

[542] Megan Sickmeier, Justin A Hamilton, Tanguy LeGall, Vladimir Vacic, Marc S Cortese, Agnes Tantos, Beata Szabo, Peter Tompa, Jake Chen, Vladimir N Uversky, Zoran

Obradovic, and A Keith Dunker. DisProt: the database of disordered proteins. *Nucleic Acids Res.*, 35:D786–D793, 2007.

[543] Paul B Sigler. Acid blobs & negative noodles. *Nature*, 333:210–212, 24 May 1988.

[544] Kresimir Sikic, Sanja Tomic, and Oliviero Carugo. Systematic comparison of crystal and NMR protein structures deposited in the protein data bank. *Open Biochem. J.*, 4:83–95, 3 September 2010.

[545] M M Silva, P H Rogers, and A Arnone. A third quaternary structure of human hemoglobin a at 1.7-a resolution. *J. Biol. Chem.*, 1992.

[546] Adelene Y L Sim, Jan Lipfert, Daniel Herschlag, and Sebastian Doniach. Salt dependence of the radius of gyration and flexibility of single-stranded DNA in solution probed by small-angle x-ray scattering. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.*, 86(2 Pt 1):021901, August 2012.

[547] Joseph R Simon, Nick J Carroll, Michael Rubinstein, Ashutosh Chilkoti, and Gabriel P López. Programming molecular self-assembly of intrinsically disordered proteins containing sequences of low complexity. *Nat. Chem.*, adance online publication, http://dx.doi.org/10.1038/nchem.2715, 30 January 2017.

[548] Charles E Sing. Development of the modern theory of polymeric complex coacervation. *Adv. Colloid Interface Sci.*, 29 April 2016.

[549] Sukrit Singh and Gregory R Bowman. Quantifying allosteric communication via both concerted structural changes and conformational disorder with CARDS. *J. Chem. Theory Comput.*, 13(4):1509–1517, 11 April 2017.

[550] John J Skinner, Wookyung Yu, Elizabeth K Gichana, Michael C Baxa, James R Hinshaw, Karl F Freed, and Tobin R Sosnick. Benchmarking all-atom simulations using hydrogen exchange. *Proc. Natl. Acad. Sci. U. S. A.*, 111(45):15975–15980, 11 November 2014.

[551] Judith E Sleeman and Laura Trinkle-Mulcahy. Nuclear bodies: new insights into assembly/dynamics and disease relevance. *Curr. Opin. Cell Biol.*, 28:76–83, 2014.

[552] Jianhui Song, Gregory-Neal Gomes, Claudiu C Gradinaru, and Hue Sun Chan. An adequate account of excluded volume is necessary to infer compactness and asphericity of disordered proteins by forster resonance energy transfer. *J. Phys. Chem. B*, 119(49):15191–15202, 10 December 2015.

[553] AK Soper, EW Castner, Jr., and Alenka Luzar. Impact of urea on water structure: a clue to its properties as a denaturant? *Biophys. Chem.*, 105(2–3):649–666, September 2003.

[554] Andrea Soranno, Brigitte Buchli, Daniel Nettels, Ryan R Cheng, Sonja Müller-Späth, Shawn H Pfeil, Armin Hoffmann, Everett A Lipman, Dmitrii E Makarov, and Benjamin Schuler. Quantifying internal friction in unfolded and intrinsically disordered proteins with single-molecule spectroscopy. *Proc. Natl. Acad. Sci. U. S. A.*, 109(44):17800–17806, 30 October 2012.

[555] Andrea Soranno, Andrea Holla, Fabian Dingfelder, Daniel Nettels, Dmitrii E Makarov, and Benjamin Schuler. Integrated view of internal friction in unfolded proteins from single-molecule FRET, contact quenching, theory, and simulations. *Proc. Natl. Acad. Sci. U. S. A.*, 114(10):E1833–E1839, 7 March 2017.

[556] Andrea Soranno, Iwo Koenig, Madeleine B Borgia, Hagen Hofmann, Franziska Zosel, Daniel Nettels, and Benjamin Schuler. Single-molecule spectroscopy reveals polymer effects of disordered proteins in crowded environments. *Proc. Natl. Acad. Sci. U. S. A.*, 111(13):4874–4879, 1 April 2014.

[557] T R Sosnick and D Barrick. The folding of single domain proteins–have we reached a consensus? *Curr. Opin. Struct. Biol.*, 21(1):12–24, 2011.

[558] David L Spector. SnapShot: Cellular bodies. *Cell*, 127(5):1071, 1 December 2006.

[559] M O Steinhauser. A molecular dynamics study on universal properties of polymer chains in different solvent qualities. part i. a review of linear chain properties. *J. Chem. Phys.*, 122(9):094901, 2005.

[560] S C Strickfaden, M J Winters, G Ben-Ari, R E Lamson, M Tyers, and P M Pryciak. A mechanism for cell-cycle regulation of MAP kinase signaling in a yeast differentiation pathway. *Cell*, 128(3):519–531, 2007.

[561] Martin C Stumpe and Helmut Grubmüller. Interaction of urea with amino acids: implications for urea-induced protein denaturation. *J. Am. Chem. Soc.*, 129(51):16126–16131, 30 November 2007.

[562] Martin C Stumpe and Helmut Grubmüller. Polar or apolar–the role of polarity for urea-induced protein denaturation. *PLoS Comput. Biol.*, 4(11):e1000221, November 2008.

[563] Xiaolei Su, Jonathon A Ditlev, Enfu Hui, Wenmin Xing, Sudeep Banjade, Julia Okrut, David S King, Jack Taunton, Michael K Rosen, and Ronald D Vale. Phase separation of signaling molecules promotes T cell receptor signal transduction. *Science*, 352(6285):595–599, 29 April 2016.

[564] Kenji Sugase, H Jane Dyson, and Peter E Wright. Mechanism of coupled folding and binding of an intrinsically disordered protein. *Nature*, 447(7147):1021–1025, 21 June 2007.

[565] Yuji Sugita and Yuko Okamoto. Replica-exchange molecular dynamics method for protein folding. *Chem. Phys. Lett.*, 314(1–2):141–151, 26 November 1999.

[566] T Suwanmajo and J Krishnan. Biphasic responses in multi-site phosphorylation systems. *J. R. Soc. Interface*, 10(89), 2013.

[567] K Suzuki, T Ehara, T Osafune, H Kuroiwa, S Kawano, and T Kuroiwa. Behavior of mitochondria, chloroplasts and their nuclei during the mitotic cycle in the ultramicroalga cyanidioschyzon merolae. *Eur. J. Cell Biol.*, 63(2):280–288, April 1994.

[568] D Svergun. Determination of the integral parameters of particles. In George W Taylor, editor, *Structure analysis by small-angle x-ray and neutron scattering*, 3, pages 59–107. Plenum, New York and London, 1987.

[569] D Svergun, C Barberato, and M H J Koch. CRYSOL - a program to evaluate x-ray solution scattering of biological macromolecules from atomic coordinates. *J. Appl. Crystallogr.*, 28:768–773, 1995.

[570] K Tamiola and F A Mulder. Using NMR chemical shifts to calculate the propensity for structural order and disorder in proteins. *Biochem. Soc. Trans.*, 40(5):1014–1020, 2012.

[571] Shiho Tanaka, Cheryl A Kerfeld, Michael R Sawaya, Fei Cai, Sabine Heinhorst, Gordon C Cannon, and Todd O Yeates. Atomic-level models of the bacterial carboxysome shell. *Science*, 319(5866):1083–1086, 22 February 2008.

[572] C Tanford. Protein denaturation. *Adv. Protein Chem.*, 23:121–282, 1968.

[573] C Tanford. Protein denaturation. c. theoretical models for the mechanism of denaturation. *Adv. Protein Chem.*, 24:1–95, 1970.

[574] C Tanford. The hydrophobic effect and the organization of living matter. *Science*, 200(4345):1012–1018, 2 June 1978.

[575] V G Taratuta, A Holschbach, G M Thurston, D Blankschtein, and G B Benedek. Liquid-liquid phase separation of aqueous lysozyme solutions: effects of ph and salt identity. *J. Phys. Chem.*, 94(5):2140–2144, 1990.

[576] Daniel P Teufel, Christopher M Johnson, Jenifer K Lum, and Hannes Neuweiler. Backbone-driven collapse in unfolded protein chains. *J. Mol. Biol.*, 409(2):250–262, 8 April 2011.

[577] T Tezuka-Kawakami, C Gell, D J Brockwell, S E Radford, and D A Smith. Urea-induced unfolding of the immunity protein im9 monitored by spFRET. *Biophys. J.*, 91(5):L42–4, 2006.

[578] F X Theillet, C Smet-Nocca, S Liokatis, R Thongwichian, J Kosten, M K Yoon, R W Kriwacki, I Landrieu, G Lippens, and P Selenko. Cell signaling, post-translational protein modifications and NMR spectroscopy. *J. Biomol. NMR*, 54(3):217–236, 2012.

[579] Francois-Xavier Theillet, Andres Binolfi, Beata Bekei, Andrea Martorana, Honor May Rose, Marchel Stuiver, Silvia Verzini, Dorothea Lorenz, Marleen van Rossum, Daniella Goldfarb, and Philipp Selenko. Structural disorder of monomeric α-synuclein persists in mammalian cells. *Nature*, 530(7588):45–50, 4 February 2016.

[580] Michael Thommen, Wolf Holtkamp, and Marina V Rodnina. Co-translational protein folding: progress and methods. *Curr. Opin. Struct. Biol.*, 42:83–89, February 2017.

[581] M Tirado-Miranda, C Haro-Perez, M Quesada-Perez, J Callejas-Fernandez, and R Hidalgo-Alvarez. Effective charges of colloidal particles obtained from collective diffusion experiments. *J. Colloid Interface Sci.*, 263(1):74–79, 2003.

[582] Maria E Tomasso, Micheal J Tarver, Deepa Devarajan, and Steven T Whitten. Hydrodynamic radii of intrinsically disordered proteins determined from experimental polyproline II propensities. *PLoS Comput. Biol.*, 12(1):e1004686, January 2016.

[583] P Tompa and P Csermely. The role of structural disorder in the function of RNA and protein chaperones. *The FASEB Journal*, 2004.

[584] Peter Tompa. Intrinsically unstructured proteins. *Trends Biochem. Sci.*, 27(10):527–533, October 2002.

[585] Peter Tompa. Intrinsically disordered proteins: a 10-year recap. *Trends Biochem. Sci.*, 37(12):509–516, 16 December 2012.

[586] Peter Tompa, Norman E Davey, Toby J Gibson, and M Madan Babu. A million peptide motifs for the molecular biologist. *Mol. Cell*, 55(2):161–169, 17 July 2014.

[587] Peter Tompa and Alan Fersht. *Structure and Function of Intrinsically Disordered Proteins*. CRC Press, 18 November 2009.

[588] Peter Tompa and Monika Fuxreiter. Fuzzy complexes: polymorphism and structural disorder in protein–protein interactions. *Trends Biochem. Sci.*, 33(1):2–8, 2008.

[589] Agnes Toth-Petroczy, Perry Palmedo, John Ingraham, Thomas A Hopf, Bonnie Berger, Chris Sander, and Debora S Marks. Structured states of disordered proteins from genomic sequences. *Cell*, 167(1):158–170.e12, 22 September 2016.

[590] Valentina Tozzini. Coarse-grained models for proteins. *Curr. Opin. Struct. Biol.*, 15(2):144–150, April 2005.

[591] Hoang T Tran, Albert Mao, and Rohit V Pappu. Role of backbone-solvent interactions in determining conformational equilibria of intrinsically disordered proteins. *J. Am. Chem. Soc.*, 130(23):7380–7392, 11 June 2008.

[592] Hoang T Tran and Rohit V Pappu. Toward an accurate theoretical framework for describing ensembles for proteins under strongly denaturing conditions. *Biophys. J.*, 91(5):1868–1886, 1 September 2006.

[593] Hoang T Tran, Xiaoling Wang, and Rohit V Pappu. Reconciling observations of sequence-specific conformational propensities with the generic polymeric behavior of denatured proteins. *Biochemistry*, 44(34):11369–11380, 30 August 2005.

[594] G Tria, H D T Mertens, M Kachala, and D I Svergun. Advanced ensemble modelling of flexible macromolecules using x-ray solution scattering. *IUCrJ*, 2:207–217, 2015.

[595] Olga G Troyanskaya, Ora Arbell, Yair Koren, Gad M Landau, and Alexander Bolshoy. Sequence complexity profiles of prokaryotic genomic sequences: a fast algorithm for calculating linguistic complexity. *Bioinformatics*, 18(5):679–688, May 2002.

[596] T Y Tsong and R L Baldwin. A sequential model of nucleation-dependent protein folding: kinetic studies of ribonuclease a. *J. Mol. Biol.*, 63(3):453–469, 14 February 1972.

[597] M J Tucker, R Oyola, and F Gai. Conformational distribution of a 14-residue peptide in solution: a fluorescence resonance energy transfer study. *J. Phys. Chem. B*, 109(10):4788–4795, 2005.

[598] Jeffrey A Ubersax and James E Ferrell, Jr. Mechanisms of specificity in protein phosphorylation (vol 8, pg 530, 2007). *Nat. Rev. Mol. Cell Biol.*, 8(8):665–665, July 2007.

[599] J B Udgaonkar. Polypeptide chain collapse and protein folding. *Arch. Biochem. Biophys.*, 531(1-2):24–33, 2013.

[600] UniProt, Consortium. UniProt: a hub for protein information. *Nucleic Acids Res.*, 43(Database issue):D204–12, 2015.

[601] Dustin Updike and Susan Strome. P granule assembly and function in caenorhabditis elegans germ cells. *J. Androl.*, 31(1):53–60, January 2010.

[602] D W Urry, B Starcher, and S M Partridge. Coacervation of solubilized elastin effects a notable conformational change. *Nature*, 222(5195):795–796, 1969.

[603] Vladimir N Uversky. Natively unfolded proteins: a point where biology waits for physics. *Protein Sci.*, 11(4):739–756, April 2002.

[604] Vladimir N Uversky. A Protein-Chameleon: Conformational plasticity of α-Synuclein, a disordered protein involved in neurodegenerative disorders. *J. Biomol. Struct. Dyn.*, 21(2):211–234, 2003.

[605] Vladimir N Uversky, Christopher J Oldfield, and A Keith Dunker. Intrinsically disordered proteins in human diseases: introducing the D2 concept. *Annu. Rev. Biophys.*, 37:215–246, 2008.

[606] Vladimir Vacic, Christopher J Oldfield, Amrita Mohan, Predrag Radivojac, Marc S Cortese, Vladimir N Uversky, and A Keith Dunker. Characterization of molecular recognition features, MoRFs, and their binding partners. *J. Proteome Res.*, 6(6):2351–2366, June 2007.

[607] B K Vainshtein. Three-dimensional electron microscopy of biological macromolecules. *Soviet Physics Uspekhi*, 16(2):185, 1973.

[608] Robin van der Lee, Marija Buljan, Benjamin Lang, Robert J Weatheritt, Gary W Daughdrill, A Keith Dunker, Monika Fuxreiter, Julian Gough, Joerg Gsponer, David T Jones, Philip M Kim, Richard W Kriwacki, Christopher J Oldfield, Rohit V Pappu, Peter Tompa, Vladimir N Uversky, Peter E Wright, and M Madan Babu. Classification of intrinsically disordered regions and proteins. *Chem. Rev.*, 114(13):6589–6631, 9 July 2014.

[609] Kim Van Roey, Bora Uyar, Robert J Weatheritt, Holger Dinkel, Markus Seiler, Aidan Budd, Toby J Gibson, and Norman E Davey. Short linear motifs: ubiquitous and functionally diverse protein interaction modules directing cell regulation. *Chem. Rev.*, 114(13):6733–6778, 9 July 2014.

[610] Arthur Veis. A review of the early development of the thermodynamics of the complex coacervation phase separation. *Adv. Colloid Interface Sci.*, 167(1-2):2–11, 14 September 2011.

[611] R Verma, R S Annan, M J Huddleston, S A Carr, G Reynard, and R J Deshaies. Phosphorylation of sic1p by G1 cdk required for its degradation and entry into S phase. *Science*, 278(5337):455–460, 1997.

[612] R Visintin and A Amon. The nucleolus: the magician's hat for cell cycle tricks. *Curr. Opin. Cell Biol.*, 12(3):372–377, 2000.

[613] Andreas Vitalis and Rohit V Pappu. ABSINTH: A new continuum solvation model for simulations of polypeptides in aqueous solutions. *J. Comput. Chem.*, 30(5):673–699, 15 April 2009.

[614] Andreas Vitalis and Rohit V. Pappu. Chapter 3 methods for monte carlo simulations of biomacromolecules. In Ralph A. Wheeler, editor, *Annual Reports in Computational Chemistry*, volume Volume 5, pages 49–76. Elsevier, 2009.

[615] Andreas Vitalis and Rohit V Pappu. A simple molecular mechanics integrator in mixed rigid body and dihedral angle space. *J. Chem. Phys.*, 141(3):034105, 21 July 2014.

[616] Andreas Vitalis, Xiaoling Wang, and Rohit V Pappu. Quantitative characterization of intrinsic disorder in polyglutamine: Insights from analysis based on polymer theories. *Biophys. J.*, 93(6):1923–1937, 15 September 2007.

[617] Vincent A Voelz, Gregory R Bowman, Kyle Beauchamp, and Vijay S Pande. Molecular simulation of ab initio protein folding for a millisecond folder NTL9(139). *J. Am. Chem. Soc.*, 132(5):1526–1528, 2010.

[618] Ekaterina Voronina, Geraldine Seydoux, Paolo Sassone-Corsi, and Ippei Nagamori. RNA granules in germ cells. *Cold Spring Harb. Perspect. Biol.*, 3(12), 2011.

[619] C Wagner. Theorie of aging of precipitates through recrystallization (ostwald ripening). *Zeitschrift Elektrochem Ber Bunsenges Phys Chem*, 65:581–591, 1961.

[620] C D Waldburger, T Jonsson, and R T Sauer. Barriers to protein folding: formation of buried polar interactions is a slow step in acquisition of structure. *Proc. Natl. Acad. Sci. U. S. A.*, 93(7):2629–2634, 2 April 1996.

[621] C D Waldburger, J F Schildbach, and R T Sauer. Are buried salt bridges important for protein stability and conformational specificity? *Nat. Struct. Biol.*, 2(2):122–128, February 1995.

[622] Edward W J Wallace, Jamie L Kear-Scott, Evgeny V Pilipenko, Michael H Schwartz, Pawel R Laskowski, Alexandra E Rojek, Christopher D Katanski, Joshua A Riback, Michael F Dion, Alexander M Franks, Edoardo M Airoldi, Tao Pan, Bogdan A Budnik, and D Allan Drummond. Reversible, specific, active aggregates of endogenous proteins assemble upon heat stress. *Cell*, 162(6):1286–1298, 10 September 2015.

[623] Jonathan Walter, Jan Sehrt, Jadran Vrabec, and Hans Hasse. Molecular dynamics and experimental study of conformation change of Poly(N-isopropylacrylamide) hydrogels in mixtures of water and methanol. *J. Phys. Chem. B*, 116(17):5251–5259, 3 May 2012.

[624] F Wang and D P Landau. Efficient, multiple-range random walk algorithm to calculate the density of states. *Phys. Rev. Lett.*, 86(10):2050–2053, 5 March 2001.

[625] Jennifer T Wang, Jarrett Smith, Bi-Chang Chen, Helen Schmidt, Dominique Rasoloson, Alexandre Paix, Bramwell G Lambrus, Deepika Calidas, Eric Betzig, and Geraldine Seydoux. Regulation of RNA granule dynamics by phosphorylation of serine-rich, intrinsically disordered proteins in c. elegans. *Elife*, 3:e04591, 23 December 2014.

[626] L Wang, J Xie, and P G Schultz. Expanding the genetic code. *Annu. Rev. Biophys. Biomol. Struct.*, 35:225–249, 2006.

[627] Xiaoling Wang, Andreas Vitalis, Matthew A Wyczalkowski, and Rohit V Pappu. Characterizing the conformational ensemble of monomeric polyglutamine. *Proteins*, 63(2):297–311, 1 May 2006.

[628] Yuanyuan Wang, Jill Trewhella, and David P Goldenberg. Small-angle x-ray scattering of reduced ribonuclease a: effects of solution conditions and comparisons with a computational model of unfolded proteins. *J. Mol. Biol.*, 377(5):1576–1592, 11 April 2008.

[629] Z Wang and J Moult. SNPs, protein structure, and disease. *Hum. Mutat.*, 17(4):263–270, April 2001.

[630] J J Ward, J S Sodhi, L J McGuffin, B F Buxton, and D T Jones. Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *J. Mol. Biol.*, 337(3):635–645, 26 March 2004.

[631] J R Warner. The nucleolus and ribosome formation. *Curr. Opin. Cell Biol.*, 2(3):521–527, 1990.

[632] G Warren and W Wickner. Organelle inheritance. *Cell*, 84(3):395–400, 9 February 1996.

[633] Stephanie C Weber and Clifford P Brangwynne. Inverse size scaling of the nucleolus by a concentration-dependent phase transition. *Curr. Biol.*, 25(5):641–646, 2 March 2015.

[634] S Weerasinghe and P E Smith. Cavity formation and preferential interactions in urea solutions: Dependence on urea aggregation. *J. Chem. Phys.*, 118(13):5901–5910, 2003.

[635] Samantha Weerasinghe and Paul E Smith. A KirkwoodâBuff derived force field for mixtures of urea and water. *J. Phys. Chem. B*, 107(16):3891–3898, 2003.

[636] Samantha Weerasinghe and Paul E Smith. A Kirkwood-Buff derived force field for the simulation of aqueous guanidinium chloride solutions. *J. Chem. Phys.*, 121(5):2180–2186, 1 August 2004.

[637] Haiyan Wei, Qiang Shao, and Yi Qin Gao. The effects of side chain hydrophobicity on the denaturation of simple [small beta]-hairpins. *Phys. Chem. Chem. Phys.*, 12(32):9292–9299, 6 August 2010.

[638] D B Wetlaufer. Nucleation, rapid folding, and globular intrachain regions in proteins. *Proc. Natl. Acad. Sci. U. S. A.*, 70(3):697–701, March 1973.

[639] Joshua R Wheeler, Tyler Matheny, Saumya Jain, Robert Abrisch, and Roy Parker. Distinct stages in stress granule assembly and disassembly. *Elife*, 5, 7 September 2016.

[640] Stephen Whitelam and Phillip L Geissler. Avoiding unphysical kinetic traps in monte carlo simulations of strongly attractive particles. *J. Chem. Phys.*, 127(15):154101, 21 October 2007.

[641] D K Wilkins, S B Grimshaw, V Receveur, C M Dobson, J A Jones, and L J Smith. Hydrodynamic radii of native and denatured proteins measured by pulse field gradient NMR techniques. *Biochemistry*, 38(50):16424–16431, 14 December 1999.

[642] S Williams, T P Causgrove, R Gilmanshin, K S Fang, R H Callender, W H Woodruff, and R B Dyer. Fast events in protein folding: helix melting and formation in a small peptide. *Biochemistry*, 35(3):691–697, 23 January 1996.

[643] M P Williamson. The structure and function of proline-rich regions in proteins. *Biochem. J*, 297 (Pt 2):249–260, 15 January 1994.

[644] Tim E Williamson, Andreas Vitalis, Scott L Crick, and Rohit V Pappu. Modulation of polyglutamine conformations and dimer formation by the n-terminus of huntingtin. *J. Mol. Biol.*, 396(5):1295–1309, 12 March 2010.

[645] E B Wilson. The structure of protoplasm. *Science*, 80(10):33–45, 1899.

[646] D S Wishart, C G Bigam, A Holm, R S Hodges, and B D Sykes. 1h, 13C and 15N random coil NMR chemical shifts of the common amino acids. i. investigations of nearest-neighbor effects. *J. Biomol. NMR*, 5(1):67–81, January 1995.

[647] D S Wishart and B D Sykes. The 13C chemical-shift index: a simple method for the identification of protein secondary structure using 13C chemical-shift data. *J. Biomol. NMR*, 4(2):171–180, March 1994.

[648] Peter G Wolynes. Evolution, energy landscapes and the paradoxes of protein folding. *Biochimie*, 119:218–230, December 2015.

[649] Jeffrey B Woodruff, Beatriz Ferreira Gomes, Per O Widlund, Julia Mahamid, and Anthony A Hyman. The centrosome is a selective phase that nucleates microtubules by concentrating tubulin. 10 December 2016.

[650] Derek N Woolfson, Gail J Bartlett, Antony J Burton, Jack W Heal, Ai Niitsu, Andrew R Thomson, and Christopher W Wood. De novo protein design: how do we expand into the universe of possible protein structures? *Curr. Opin. Struct. Biol.*, 33:16–26, August 2015.

[651] John C Wootton and Scott Federhen. Statistics of local complexity in amino acid sequences and sequence databases. *Comput. Chem.*, 17(2):149–163, June 1993.

[652] P E Wright and H J Dyson. Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *J. Mol. Biol.*, 293(2):321–331, 22 October 1999.

[653] Peter E Wright and H Jane Dyson. Linking folding and binding. *Curr. Opin. Struct. Biol.*, 19(1):31–38, February 2009.

[654] Peter E Wright and H Jane Dyson. Intrinsically disordered proteins in cellular signalling and regulation. *Nat. Rev. Mol. Cell Biol.*, 16(1):18–29, 22 December 2014.

[655] Y Wu, E Kondrashkina, C Kayatekin, C R Matthews, and O Bilsel. Microsecond acquisition of heterogeneous structure in the folding of a TIM barrel protein. *Proc. Natl. Acad. Sci. U. S. A.*, 105(36):13367–13372, 2008.

[656] Matthew A Wyczalkowski and Rohit V Pappu. Satisfying the fluctuation theorem in free-energy calculations with hamiltonian replica exchange. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.*, 77(2 Pt 2):026104, 11 February 2008.

[657] Junchao Xia, William F Flynn, Emilio Gallicchio, Bin W Zhang, Peng He, Zhiqiang Tan, and Ronald M Levy. Large-scale asynchronous and distributed multidimensional replica exchange molecular simulations and efficiency analysis. *J. Comput. Chem.*, 36(23):1772–1785, 5 September 2015.

[658] Zhen Xia, Payel Das, Eugene I Shakhnovich, and Ruhong Zhou. Collapse of unfolded proteins in a mixture of denaturants. *J. Am. Chem. Soc.*, 134(44):18266–18274, 2012.

[659] Siheng Xiang, Masato Kato, Leeju C Wu, Yi Lin, Ming Ding, Yajie Zhang, Yonghao Yu, and Steven L McKnight. The LC domain of hnRNPA2 adopts similar conformations in hydrogel polymers, liquid-like droplets, and nuclei. *Cell*, 163(4):829–839, 11 May 2015.

[660] Koji Yamano and Richard J Youle. Coupling mitochondrial and cell division. *Nat. Cell Biol.*, 13(9):1026–1027, 2 September 2011.

[661] Tae Yeon Yoo, Steve P Meisburger, James Hinshaw, Lois Pollack, Gilad Haran, Tobin R Sosnick, and Kevin Plaxco. Small-Angle x-ray scattering and Single-Molecule FRET spectroscopy produce highly divergent views of the Low-Denaturant unfolded state. *J. Mol. Biol.*, 418(3-4):226–236, 27 January 2012.

[662] A Yu. Grosberg, S K Nechaev, and E I Shakhnovich. The role of topological constraints in the kinetics of collapse of macromolecules. *Journal de Physique*, 49(12):2095–2100, 1 December 1988.

[663] T Yuwen and N R Skrynnikov. CP-HISQC: a better version of HSQC experiment for intrinsically disordered proteins under physiological conditions. *J. Biomol. NMR*, 58(3):175–192, 2014.

[664] Taraneh Zarin, Caressa N Tsai, Alex N Nguyen Ba, and Alan M Moses. Selection maintains signaling function of a highly diverged intrinsically disordered region. *Proc. Natl. Acad. Sci. U. S. A.*, 114(8):E1450–E1459, 21 February 2017.

[665] Gül H Zerze, Robert B Best, and Jeetain Mittal. Modest influence of FRET chromophores on the properties of unfolded proteins. *Biophys. J.*, 107(7):1654–1660, 7 October 2014.

[666] Gül H Zerze, Robert B Best, and Jeetain Mittal. Sequence- and Temperature-Dependent properties of unfolded and disordered proteins from atomistic simulations. *J. Phys. Chem. B*, 119(46):14622–14630, 19 November 2015.

[667] Gül H Zerze, Cayla M Miller, Daniele Granata, and Jeetain Mittal. Free energy surface of an intrinsically disordered protein: comparison between temperature replica exchange molecular dynamics and bias-exchange metadynamics. *J. Chem. Theory Comput.*, 11(6):2776–2782, 9 June 2015.

[668] Huaiying Zhang, Shana Elbaum-Garfinkle, Erin M Langdon, Nicole Taylor, Patricia Occhipinti, Andrew A Bridges, Clifford P Brangwynne, and Amy S Gladfelter. RNA controls PolyQ protein phase transitions. *Mol. Cell*, 60(2):220–230, 15 October 2015.

[669] Jing Zhang, Xiangdong Peng, Ana Jonas, and Jiri Jonas. NMR study of the cold, heat, and pressure unfolding of ribonuclease a. *Biochemistry*, 34(27):8631–8641, 1995.

[670] Yi Zhang, Alejandro Wolf-Yadlin, Phillip L Ross, Darryl J Pappin, John Rush, Douglas A Lauffenburger, and Forest M White. Time-resolved mass spectrometry of tyrosine phosphorylation sites in the epidermal growth factor receptor signaling network reveals dynamic modules. *Mol. Cell. Proteomics*, 4(9):1240–1250, September 2005.

[671] R Zhao, M S Bodnar, and D L Spector. Nuclear neighborhoods and gene expression. *Curr. Opin. Genet. Dev.*, 19(2):172–179, 2009.

[672] Wenwei Zheng, Alessandro Borgia, Madeleine B Borgia, Benjamin Schuler, and Robert B Best. Empirical optimization of interactions between proteins and chemical denaturants in molecular simulations. *J. Chem. Theory Comput.*, 11(11):5543–5553, 10 November 2015.

[673] Wenwei Zheng, Alessandro Borgia, Karin Buholzer, Alexander Grishaev, Benjamin Schuler, and Robert B Best. Probing the action of chemical denaturant on an intrinsically disordered protein by simulation and experiment. *J. Am. Chem. Soc.*, 138(36):11702–11713, 14 September 2016.

[674] Dongqiang Zhu, Bruce E Herbert, Mark A Schlautman, and Elizabeth R Carraway. Characterization of cation-pi interactions in aqueous solution using deuterium nuclear magnetic resonance spectroscopy. *J. Environ. Qual.*, 33(1):276–284, January 2004.

[675] Bruno H Zimm. Dynamics of polymer molecules in dilute solution: Viscoelasticity, flow birefringence and dielectric loss. *J. Chem. Phys.*, 24(2):269–278, 1 February 1956.

[676] Maxwell I Zimmerman and Gregory R Bowman. FAST conformational searches by balancing Exploration/Exploitation Trade-Offs. *J. Chem. Theory Comput.*, 11(12):5747–5757, 2015.

[677] Ewa Zlotek-Zlotkiewicz, Sylvain Monnier, Giovanni Cappello, Mael Le Berre, and Matthieu Piel. Optical volume and mass measurements show that mammalian cells swell during mitosis. *J. Cell Biol.*, 211(4):765–774, 23 November 2015.

[678] David Zwicker, Anthony A Hyman, and Frank Jülicher. Suppression of ostwald ripening in active emulsions. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.*, 92(1):012317, July 2015.